Detecting Convoys Using License Plate Recognition Data

Sean Lawlor, *Student Member, IEEE*, Timothy Sider, Naveen Eluru, Marianne Hatzopoulou, and Michael G. Rabbat, *Senior Member, IEEE*

***Abstract*—*License plate recognition* (LPR) sensors are embedded camera systems that monitor road traffic. When a vehicle passes by a sensor, the vehicle's license plate, the location, and the time of observation are recorded. Given a stream of such observations from a collection of sensors spread around the road network, our goal is to detect convoys: groups of two or more vehicles traveling with highly correlated trajectories. Some of the main challenges with modeling and processing data from LPR sensors include that the data-gathering process is event-driven, thus data are not regularly sampled in time or space. Also, an appropriate definition of convoy should be relative to background traffic patterns which are temporally and spatially varying. This paper proposes novel models for LPR observations of traffic which are well-suited for online convoy detection. Baseline traffic is modeled as following a mixture of semi-Markov processes, and specific models for temporal and spatial correlation of observations of vehicles traveling in a convoy are introduced. These models are used within a sequential hypothesis testing framework to obtain a system for real-time convoy detection. The model of baseline traffic may be of independent interest for forecasting road traffic patterns. Experiments with an extensive simulated dataset illustrate the performance of the scheme and offer insights into the tradeoffs between detection rate, false alarm rate, and the expected number of observations required to detect a convoy.**

I. Introduction

We consider the problem of detecting convoys of vehicles in an urban environment using a collection of *license plate recognition* (LPR) sensors. Each sensor records data of the form "vehicle X was observed at location Y at time t". Given streams of such observations arriving from a collection of sensors, a centralized decision maker must identify which, if any, vehicles are traveling as convoys.

Convoy detection has applications in both law enforcement and the commercial sector. Law enforcement agents may be interested in detecting and tracking convoys for a variety of reasons [2]. In the commercial sector, the approach developed in this paper could be used to identify groups of shipping vehicles traveling along highly correlated routes which may benefit from forming platoons. Recently there has been interest in designing control laws to allow heavy-duty shipping vehicles to maintain platoons over long distances in order to reduce drag on the non-leader vehicles, thereby saving on fuel costs [3], [4]. In order to exploit this approach one must first identify potential pairs of vehicles that could form platoons, and the convoy detection approach we propose could be used to automate this process.

Defining a concrete notion of what it means to be a convoy is not as straightforward as it may seem. Intuitively a convoy comprises two or more vehicles traveling together. While it may be tempting to particularize this definition to say that a convoy is two or more vehicles traveling along the same route over a given distance (e.g., for more than 500 consecutive meters) or for a minimum amount of time (e.g., at least 5 minutes), without separating by more than a particular distance (e.g., 50 meters), such a threshold-based approach has a number of limitations and drawbacks. Setting the thresholds too tight does not allow for situations where the convoy vehicles take slightly different routes (e.g., deviating for a few city blocks before rejoining). Similarly, in dense urban environments or along stretches of highway during rush hour it may be expected that arbitrary vehicles will be seen near each other for a relatively long distance and/or time even if they are not traveling as a convoy, simply because of the dense traffic.

Similar to problems of unsupervised novelty/anomaly detection [5]–[7], defining what it means to be a convoy is not straightforward. One may expect convoys to be relatively rare events. Still it is not straightforward to obtain a sample of traffic that is guaranteed to contain no convoys, and it is also not straightforward to obtain labeled examples of convoys for training. Intuitively, two vehicles may be called a convoy if their trajectories are more correlated in space and time than two typical vehicles in normal traffic. The challenge is in making precise what is "more correlated" and what are "typical vehicles in normal traffic".

Another challenge is due to the fact that measurements arrive at irregular times. Existing LPR sensors use cameras in conjunction with computer vision algorithms to identify and extract vehicle license plates. Consequently, LPR sensors have a short range, and measurements are obtained in an event-driven manner, when a vehicle passes within the field of view

This work was funded by Genetec Inc., and by the Natural Sciences and Engineering Research Council of Canada via grant CRDPJ 446413-12.

S. Lawlor and M.G. Rabbat are with the Department of Electrical and Computer Engineering, McGill University, Montreal, Quebec, H3A 0E9, Canada. T. Sider is a Transportation Modeler with Atkins, London, NW1 3AT, United Kingdom. N. Eluru is with the Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, 32816. M. Hatzopoulou is with the Department of Civil Engineering, McGill University, Montreal, Quebec, H3A 0E9, Canada. E-mail: sean.lawlor@mail.mcgill.ca, michael.rabbat@mcgill.ca, timothy.sider@mail.mcgill.ca, marianne.hatzopoulou@mcgill.ca, naveen.eluru@ucf.edu.

A preliminary version of this work was presented at the 2014 Asilomar Conference on Signals, Systems, and Computers [1].

of the camera. Thus, measurement times are arbitrary, and measurements of any particular vehicle are not obtained at regular sampling intervals, either in time or space.

The aim of this work is to develop algorithms that detect convoys in real-time. Our approach is based on sequential hypothesis testing [8], and the main contribution of this work is the modeling of observations from such a network of LPR sensors. Under the independent (non-convoy) hypothesis, vehicle movement is modeled as following a mixture of Markov chains, and under the convoy hypothesis a novel leader/follower observation model is developed.

A. Previous work

The majority of previous work on convoy detection and tracking in the information fusion literature [9]–[11] focuses on sensors with a wide field of view, such as *ground moving target indicator* radar. Data is collected from one or a few sensors and provides a tracking indicator based on the physical characteristics of a vehicle. Each sensor regularly scans and gathers measurements about the vehicles in its field of view over an extended period of time and over a large geographic region. In contrast, the setting considered in this paper is such that any individual sensor only measures a vehicle when it is nearby the sensor, and individual vehicles are thus only measured intermittently (and irregularly) over time when they pass by a sensor.

Threshold-based approaches have been studied for off-line identification of convoys in trajectory databases [12]–[15]. Such methods are applicable when entire vehicle trajectories are available (e.g., all vehicles carry GPS units and regularly report their location to a central office, as is commonly the case with taxis and shipping trucks). In such a setting, when a vehicle is also aware of which other vehicles are nearby, convoys can be detected using decentralized methods [16]. In contrast to this previous work, the present paper deals with partially-observed trajectories, sampled when the vehicle passes by an LPR sensor. In addition, the previous work mentioned above does not take into account the underlying traffic patterns and structure of the road network.

Convoy detection based on LPR sensors (a.k.a., automatic number plate recognition systems) is considered in [2], where a heuristic approach to detecting vehicle convoys in a database of LPR records is proposed. The approach, similar to [13], is based on counting and thresholding co-occurrences of vehicles observed nearby each other. The convoy model considered in [2] requires that the vehicles in a convoy follow precisely the same path, and the method is designed for post-processing of database records rather than real-time/sequential detection.

B. Contributions and organization

We address the problem of convoy detection using tools from the statistical signal processing toolbox. Specifically, the contributions of this work are: 1) posing the problem of convoy detection using short-range LPR sensors in the framework of sequential hypothesis testing, and 2) developing models for LPR observations under convoy and non-convoy hypotheses. In the non-convoy setting we model vehicle movement using a mixture of Markov models. In the convoy setting, a novel leader/follower measurement model is developed. The convoy model is flexible and does not require all vehicles in the convoy to travel along precisely the same route; rather they should travel in the same general direction (following the leader), and the leader may change over time. The extent to which their routes deviate can be specified in the model, so that the scenario where all convoy vehicles follow precisely the same route is a special case. We evaluate the performance of the proposed approach using simulated data based on a detailed model of road traffic in Montréal, Canada.

The rest of the paper is organized as follows. Section II provides the problem formulation. Generative models for observations of convoys and independent vehicles are described in Section III. The proposed sequential hypothesis testing framework, including implementation details, is described in Section IV. The results of the experimental performance evaluation are reported in Section V. Additional issues are discussed in Section VI, and we conclude in Section VII.

II. Problem Description

This section takes steps towards formalizing the problem of convoy detection. We describe characteristics of the measurement system that make the problem challenging. Then we discuss assumptions made and describe performance metrics that will be used to evaluate convoy detection methods.

A. License plate recognition data

Consider a system of urban roads instrumented with license plate recognition sensors. When a vehicle passes by the sensor it records the license plate as well as the time and location of the event. The sensors have a very short range of detection (e.g., 10 meters). The measurements from many of these sensors, at different locations in the road network, report their measurements to a fusion center whose goal is to detect groups of vehicles that are driving together as a convoy.

Formally, we consider a collection of C sensors, indexed using the first C natural numbers, $1, \dots, C$, and let the set of sensor indices be $\Omega = \{1, \dots, C\}$. In this paper we focus on detecting convoys composed of two vehicles; the extension to convoys of more than two vehicles is discussed in Section VI. Let $(x_1, r_1), (x_2, r_2), \dots,$ denote a sequence of observations of one vehicle, where $x_i \in \Omega$ is the index of the sensor making the ith observation and $r_i \in \mathbb{R}_+$ is the time of the ith observation.[1] Similarly, let $(y_1, s_1), (y_2, s_2), \dots,$ denote the sequence of observations of a second vehicle.

B. Measurement system characteristics and assumptions

Fig. 1 shows a sample path of the measurement process as a function of time. The horizontal axis corresponds to time; the labels along the bottom of the figure show observation times for each vehicle (r_i and s_i), and the labels along the top of the figure show global observation times. The vertical axis gives

[1] Without loss of generality we take all times to be non-negative and denote by $\mathbb{R}_+$ the set of non-negative real numbers.

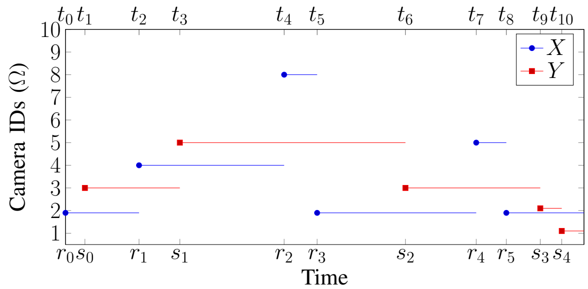


Fig. 1. Example measurements of two vehicles over time.

the index of the camera making the measurement; this index should be treated as a categorical variable since the ordering is arbitrary and does not necessarily reflect, e.g., the geography of the sensors.

Note that the observation times are not necessarily equally spaced, and the number of observations is not necessarily the same for each vehicle. This is because vehicles can leave the observation area, people drive at different speeds, and traffic patterns and road conditions vary over time.

Traffic patterns in a large urban environment may also be quite complex. For example, during rush hour there may be a significant flow of vehicles heading from the suburbs into the city and, at the same time, from the city out to the suburbs. This motivates the need for models that can capture these subtle aspects of traffic flows and not just the average or majority flow over the network.

We make the following assumptions about the measurements. First, the sensors are synchronized so that the timestamps from different sensors are directly comparable. This is justified since existing LPR cameras are typically equipped with GPS receivers that can provide reliable and accurate synchronization.

Second, we assume that no two vehicle observations are recorded at precisely the same time instant; this ensures that the two time sequences $\{r_i\}_{i\geq 1}$ and $\{s_i\}_{i\geq 1}$ can be uniquely ordered. This is justified when timestamps at each sensor use a sufficiently high resolution.

Third, we assume that a sensor always records vehicles that pass by the road segment it is monitoring and that the sensor does not produce any spurious measurements. Thus, there is no "noise" in the measurement sequences (missed observations or erroneously injected observations), and the main source of uncertainty is in the vehicle trajectories. While it is certainly of interest to allow for such additional noise sources, we leave this as an extension for future work.

Fourth, we assume that the sensors are static and that their locations are known to the fusion center. Thus, the fusion center can make use of related information, such as the distance between sensors, when making a decision.

Finally, we assume that the sensors transmit their measurements to the fusion center over a reliable, delay-free channel; i.e., we consider a traditional centralized decision making setup. This is reasonable since each individual measurement can be encoded in a small number of bits (e.g., much smaller than the size of a typical Ethernet packet) and the inter-observation time for a given vehicle (i.e., the time between when it is observed at one sensor and next observed at a different sensor) is large relative to the time it takes to transmit such a measurement using contemporary communication technologies.

C. *Sequential testing and performance metrics*

In this work we consider a typical sequential hypothesis testing setting [8] where the observations (x_i, r_i) and (y_i, s_i) arrive successively at the fusion center. Under the null hypothesis, H_0, the vehicles are independent (not a convoy), and under the alternative hypothesis, H_1, the vehicles are moving as a convoy. After receiving an observation the decision maker must choose from one of three options: 1) declare that the pair of vehicles is a convoy (i.e., reject the null), 2) declare that the pair is not a convoy (i.e., fail to reject the null), or 3) wait to receive additional observations. As discussed in Section I, defining what it means to be a convoy is difficult. Ultimately, the precise definition of convoy adopted in this work is implicit in the models described in Section III.

The objective is to make accurate decisions without deferring for too long. Accuracy is measured using the standard metrics for hypothesis testing: the probability of detection and probability of false alarm. We also study the average number of observations required to make a decision. Ideally a system should have high probability of detection, low probability of false alarm, and a low average number of observations required to make a decision.

III. Modeling

Our aim is to formulate the problem of convoy detection in the sequential hypothesis testing framework. The main task is one of modeling; i.e., to define appropriate distributions for the observations under the hypotheses that (H_1) the two observed vehicles are a convoy, or (H_0) the vehicles are not a convoy. First we describe a simple Markov model for the observations of individual vehicles. Then we build on this to develop models for observations of pairs of vehicles under each hypothesis.

A. *Single-vehicle Markov model*

To begin, we define a model for the observations of a single vehicle, $\{(x_i, r_i)\}_{i=1}^n$. Our model can be viewed as a semi-Markov process [17], where the sequence of sensors where the vehicle is observed, $x_1, x_2, \dots$, follows a discrete-time Markov chain, and the inter-observation times $r_i - r_{i-1}$, $i = 2, \dots, n$, are mutually independent and are conditionally independent of the other variables given the states x_{i-1} and x_i.[2]

Let $(\pi_x)_{x\in\Omega}$ denote the initial state distribution, with

$$\sum_{x\in\Omega} \pi_x = 1,$$

[2] If the inter-observation times were assumed to follow an exponential distribution then the semi-Markov process is equivalent to a continuous-time Markov chain. In general, the inter-observation times of a semi-Markov process may follow an arbitrary distribution with support on the positive real numbers.

and let $P_{x_{i-1},x_i} = \Pr(x_i|x_{i-1})$ denote the transition distribution of a Markov chain, satisfying

$$\sum_{x_i \in \Omega} P_{x_{i-1},x_i} = 1, \qquad \forall x_{i-1} \in \Omega \ .$$

Furthermore, let $f(r_i - r_{i-1}|x_{i-1}, x_i)$ denote the density of the ith inter-observation time given that a vehicle was observed at sensor x_{i-1} and then at x_i. We require that $f(\cdot|x_{i-1}, x_i)$ has support on $\mathbb{R}_+$ for all $x_{i-1}, x_i \in \Omega$.

Under a semi-Markov model, the likelihood of the observations $\{(x_i, r_i)\}_{i=1}^n$ is

$$\mathrm{p}(\{(x_i, r_i)\}_{i=1}^n) = \pi_{x_1} \prod_{i=2}^n P_{x_{i-1},x_i} f(r_i - r_{i-1}|x_{i-1}, x_i) \ .$$

To capture richer, more complicated traffic patterns, we modify the model on the sequence of sensors $x_1, \ldots, x_n$ which observe the vehicle to be a mixture of Markov chains. Let M be a positive integer. For $m = 1, \ldots, M$, let $\pi_x^{(m)}$ denote the initial state distribution of the mth mixture component and let $P_{x_{i-1},x_i}^{(m)}$ denote the transition probabilities of the mth component. Also let $\theta^{(1)}, \ldots, \theta^{(M)}$ be the mixture parameters, satisfying $\theta_m \geq 0$ for all $m = 1, \ldots, M$ and $\sum_{m=1}^M \theta^{(m)} = 1$.

We associate a latent variable m with each vehicle, taking values in the set $\{1, \ldots, M\}$, indicating which mixture component governs the vehicle's path. The trajectory of any particular vehicle is governed by only one component of the mixture model; i.e., each vehicle is a realization of this process and the particular component governing its trajectory is a multinomial random variable with parameters $\theta^{(1)}, \ldots, \theta^{(M)}$. Then the likelihood of the observations $\{(x_i, r_i)\}_{i=1}^n$ in the mixture model is given by

$$\begin{aligned} &\mathrm{p}(\{(x_i, r_i)\}_{i=1}^n) \\ &= \left(\sum_{m=1}^M \theta^{(m)} \pi_{x_1}^{(m)} \prod_{i=2}^n P_{x_{i-1},x_i}^{(m)} \right) \prod_{i=2}^n f(r_i - r_{i-1}|x_{i-1}, x_i) \ . \end{aligned} \tag{1}$$

Note that the mixture model only applies to the sequence of states, and the conditional distribution of the inter-observation times $r_i - r_{i-1}$ given the states x_{i-1} and x_i are independent of the mixture component m. In a transportation network this implies that the time to travel from x_{i-1} to x_i is independent of the process determining the route the vehicle is following.

In order to evaluate the likelihood (1) given observations $\{(x_i, r_i)\}_{i \geq 1}$ we need to specify the form of the inter-observation time density and we need to provide values for the parameters $\{\theta^{(m)}, \pi_x^{(m)}, P_{x,x'}^{(m)} : x, x' \in \Omega, m = 1, \ldots, M\}$ of the Markov chain mixture model. As mentioned above, the inter-observation time density $f(r_i - r_{i-1}|x_{i-1}, x_i)$ can be any density with support on the positive real numbers. Examples of potential choices include the truncated Gaussian, inverse-Gaussian, and gamma distributions. Each of these distributions has additional parameters which would need to be fit from data. In practice, we fit these parameters and the parameters of the Markov chain mixture model using data from a training period taken before the sequential hypothesis test for convoys goes online. We describe this training procedure in more detail in Section IV.

B. Notation for observations of two vehicles

Recall that, in the convoy detection problem, we have two observation sequences $\{(x_i, r_i)\}_{i \geq 1}$ and $\{(y_i, s_i)\}_{i \geq 1}$ of the two vehicles, which we will refer to as X and Y, where $x_i \in \Omega$ is the identifier of the sensor that observes vehicle X at time r_i, and where $\Omega = \{1, \ldots, C\}$ denotes the collection of sensor indices. Also recall that the times r_i and s_i do not coincide; i.e., the observation times are not regularly sampled. Towards developing models and a sequential hypothesis test involving this data, we introduce notation to allow for simultaneously indexing the observations of both vehicles.

For a given time $t \in \mathbb{R}_+$, let

$$n_x(t) = \max\{i \colon r_i \leq t\}$$

denote the number of observations of vehicle X that have been collected at time t, let

$$n_y(t) = \max\{i \colon s_i \leq t\}$$

denote the number of observations of vehicle Y that have been collected at time t, and let

$$n(t) = n_x(t) + n_y(t)$$

denote the total number of observations of either vehicle that have been collected at time t. Let

$$T(t) = \{r_i\}_{i=1}^{n_x(t)} \cup \{s_i\}_{i=1}^{n_y(t)}$$

denote the set of all times when either vehicle is observed. Based on the assumption that no two observation events occur simultaneously, the cardinality of $T(t)$ is $n(t)$ and we can write

$$T(t) = \{t_0, t_2, \ldots, t_{n(t)-1}\},$$

where $t_k < t_{k+1}$, $k = 1, \ldots, n(t)-1$; i.e., $T(t)$ can be viewed as the sequence of observation event times.

We assume that one of the two cases,

$$\{t_0 = r_1 \text{ and } t_1 = s_1\} \text{ or } \{t_0 = s_1 \text{ and } t_1 = r_1\} \ ,$$

holds; i.e., the test begins with one observation of each vehicle.

At each observation time t_k, exactly one vehicle is observed. It will be useful to define the extended observation sequences,

$$\begin{aligned} \mathbf{x}_k &= (\widetilde{x}_k, \widetilde{r}_k) \in \Omega \times \mathbb{R}_+ \\ \mathbf{y}_k &= (\widetilde{y}_k, \widetilde{s}_k) \in \Omega \times \mathbb{R}_+ \ , \end{aligned}$$

for $k = 1, \ldots, n(t) - 1$, where

$$\widetilde{x}_k = x_{n_x(t_k)} \quad \text{and} \quad \widetilde{r}_k = r_{n_x(t_k)}$$

are the sensor and time where vehicle X was last seen as of observation time t_k, and

$$\widetilde{y}_k = y_{n_y(t_k)} \quad \text{and} \quad \widetilde{s}_k = s_{n_y(t_k)}$$

are the sensor and time where vehicle Y was last seen as of observation time t_k. For example, t_k is a time when vehicle X is observed then $\widetilde{r}_k = t_k$, and $\widetilde{s}_k$ $(< \widetilde{r}_k)$ is the most recent time prior to t_k when vehicle Y is observed. We define the extended observation starting only from time t_1 (not t_0) so that both vehicles have been observed. Finally, let $\mathbf{x}_{1:n} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ denote the X-observation sequence at the first n joint observation times, and let $\mathbf{y}_{1:n}$ be defined in a similar manner.

Note that there is an equivalence between the extended observation sequence $(\mathbf{x}_{1:n(t)-1}, \mathbf{y}_{1:n(t)-1})$ and the per-vehicle observation sequences, $\{(x_i, r_i)\}_{i=1}^{n_x(t)}$ and $\{(y_i, s_i)\}_{i=1}^{n_y(t)}$, in the sense that one can always construct the extended observation sequence given the per-vehicle observation sequences, and the per-vehicle sequences can be uniquely extracted from the extended observation sequence. Hence, the two representations convey precisely the same information.

C. Two-vehicle likelihood factorization

We assume that under both of the hypotheses, H_j with $j \in \{0, 1\}$, the joint likelihood of the extended observation sequences $\mathbf{x}_{1:k}$ and $\mathbf{y}_{1:k}$ is first-order Markov; i.e.,

$$\begin{aligned} &\mathrm{p}(\mathbf{x}_{1:n(t)-1}, \mathbf{y}_{1:n(t)-1} | H_j) \\ &\quad = \pi(\mathbf{x}_1, \mathbf{y}_1) \prod_{k=2}^{n(t)-1} \mathrm{p}(\mathbf{x}_k, \mathbf{y}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, H_j) \ . \end{aligned} \tag{2}$$

This makes it possible to recursively calculate the log-likelihood ratio, simplifying the implementation of the sequential hypothesis test which is discussed further in Section IV. We also assume that the initial distribution $\pi(\mathbf{x}_1, \mathbf{y}_1)$ is independent of the hypothesis. In the following subsections we describe the proposed transition distribution $\mathrm{p}(\mathbf{x}_k, \mathbf{y}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, H_j)$ under each hypothesis $j \in \{0, 1\}$.

To simplify the notation, in the sequel we write $\mathrm{p}_j(\mathbf{x}_k, \mathbf{y}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1})$ for the transition dynamics under hypothesis $j \in \{0, 1\}$.

D. Model for vehicles traveling independently (H_0)

The null hypothesis (H_0) states that the two vehicles are traveling through the network independent of each other. The likelihood of the observed paths of the two vehicles under this null hypothesis is simply the product of the two individual likelihoods from the previous section,

$$\begin{aligned} &\mathrm{p}_0(\mathbf{x}_{1:n(t)-1}, \mathbf{y}_{1:n(t)-1}) \\ &\quad = \mathrm{p}_0\left(\{(x_i, r_i)\}_{i=1}^{n_x(t)}, \{(y_i, s_i)\}_{i=1}^{n_y(t)}\right) \\ &\quad = \mathrm{p}\left(\{(x_i, r_i)\}_{i=1}^{n_x(t)}\right) \mathrm{p}\left(\{(y_i, s_i)\}_{i=1}^{n_y(t)}\right), \end{aligned}$$

where the individual likelihood of each vehicle is given by (1).

E. Markov model for convoys (H_1)

As discussed in the introduction, giving a precise definition of a convoy is not straightforward. We seek a method where the notion of a convoy encompasses the following elements:

1) At any point in time, one vehicle is following the other, and which vehicle is leading a convoy may change at any point in time.
2) The vehicles in a convoy need not take precisely the same route, but they should remain near each other (e.g., within a prescribed distance threshold).
3) The distance between the vehicles in a convoy is roughly proportional to the speed at which they are traveling, so if the vehicles were to follow exactly the same route then the time between consecutive observations of each vehicle at the same sensor would be roughly constant.

Initially a pair of vehicles is observed close together (possibly by the same camera or a nearby camera, and near in time) in order to trigger the initialization of a hypothesis test. Then the subsequent observations of the pair can be used to update the likelihood of the convoy.

Consider two vehicles, X and Y, moving through the network as a convoy. Initial observations for both vehicles will be set to the same likelihood under H_1 as under H_0. Specifically, we take the initial state distribution to be equal under both hypotheses. This is

$$\mathrm{p}_j(\mathbf{x}_1, \mathbf{y}_1) = \max_m \left\{ \pi_{x_1}^{(m)} \pi_{y_1}^{(m)} \right\}, j \in \{0, 1\}$$

which selects the maximum likelihood mixture component for the initial distribution for the vehicles.

Under the convoy hypothesis, H_1, we model the sequence of states where the two vehicles are observed as being generated by the same mixture component in the mixture of Markov chain model. Given that this is mixture component m, the likelihood is given by

$$\begin{aligned} &\mathrm{p}_1(\mathbf{x}_{1:k}, \mathbf{y}_{1:k} | m) \\ &\quad = \pi_m(x_1) \pi_m(y_1) \times \prod_{i=2}^{n(t_k)-1} \mathrm{p}_1(\mathbf{x}_i, \mathbf{y}_i | \mathbf{x}_{i-1}, \mathbf{y}_{i-1}, m). \end{aligned}$$

Due to the assumption that exactly one observation is made at any time instant, at any observation time t_k either the observation is of X, in which case $\widetilde{r}_k > \widetilde{s}_k$, or the observation is of Y, in which case $\widetilde{s}_k > \widetilde{r}_k$. Note that if the observation at time t_k was of X (respectively, of Y), then $\mathbf{y}_k = \mathbf{y}_{k-1}$ (respectively, $\mathbf{x}_k = \mathbf{x}_{k-1}$), and thus

$$\begin{aligned} &\mathrm{p}_1(\mathbf{x}_k, \mathbf{y}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, m) \\ &\quad = \begin{cases} \mathrm{p}_1(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, m) & \text{if } \widetilde{r}_k > \widetilde{s}_k \\ \mathrm{p}_1(\mathbf{y}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, m) & \text{if } \widetilde{s}_k > \widetilde{r}_k, \end{cases} \end{aligned} \tag{3}$$

and so we must define the likelihood function for the two cases in (3).

Based on the description of a convoy given in Section II, we desire a model where, under H_1, one vehicle will be leading and the other will be following at any point in time. If, for example, X is leading, then X transitions first and Y transitions to a state afterwards which depends on X's new location. We do not require that the same vehicle lead the entire time. The leader can switch shortly after X and Y have been observed near each other; the idea is that to switch between leading and following roles the follower must pass the leader.

Let $\mathrm{dist}(x, x')$ denote the geographic distance between two sensors $x, x' \in \Omega$, and let $L > 0$ be a given proximity threshold. To capture the different possible observation scenarios under this model, we split the description of the transition distribution (3) into six cases:

1) X and Y are close at time t_{k-1} (no clear leader) and X is observed next;
2) X and Y are close at time t_{k-1} (no clear leader) and Y is observed next;
3) X is leading and X is observed next;
4) Y is leading and Y is observed next;
5) X is leading and Y is observed next;
6) Y is leading and X is observed next.

Which case applies at a given point in time can be determined by examining the following three quantities:

i) The distance between the last observations of X and Y, $\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1})$, relative to the threshold L;

ii) Which vehicle was observed at time t_{k-1},

$$\begin{cases} X & \text{if } \widetilde{r}_{k-1} > \widetilde{s}_{k-1}, \\ Y & \text{if } \widetilde{s}_{k-1} > \widetilde{r}_{k-1}; \end{cases}$$

iii) Which vehicle was observed at time t_k,

$$\begin{cases} X & \text{if } \widetilde{r}_{k} > \widetilde{s}_{k}, \\ Y & \text{if } \widetilde{s}_{k} > \widetilde{r}_{k}. \end{cases}$$

If X and Y were last observed close together (i.e., $\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1}) < L$) then there is no clear leader, so the next vehicle to be observed may do so independently of the last observed location of Y. For example, if the vehicles were seen near each other at time t_{k-1} (i.e., the distance between the two observing cameras is less than the proximity threshold L), then if the observation at time t_k is of X and the vehicles are following mixture component m we take

$$\mathrm{p}_1(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, m) = P^{(m)}_{\widetilde{x}_{k-1}, \widetilde{x}_k} f(\widetilde{r}_k - \widetilde{r}_{k-1} | \widetilde{x}_{k-1}, \widetilde{x}_k),$$

where $P^{(m)}_{x,x'}$ and $f(\cdot|x, x')$ denote the same transition and inter-observation time distributions used in the model under H_0.

If $\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1}) \geq L$, then X and Y were not last seen close together, and one of the two vehicles is leading. If the previous observation was of X (i.e., $\widetilde{r}_{k-1} > \widetilde{s}_{k-1}$) then X is leading, and if the previous observation was of Y then Y is leading. In this case we further check whether the most recent observation, at time t_k, was of the leader or of the follower.

It can happen that the leader vehicle is observed multiple times between consecutive observations of the follower. For example, if X is leading, X and Y are already separated by a distance larger than L, and X is observed again, then X is moving further away from Y. This is the case if $\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1}) \geq L$ and $\widetilde{r}_{k-1} > \widetilde{s}_k$ (i.e., X is observed again before Y, so we have at least two observations of X since the last observation of Y). In this scenario we again model X's transition as being independent of the last observation of Y,

$$\mathrm{p}_1(\mathbf{x}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, m) = P^{(m)}_{\widetilde{x}_{k-1}, \widetilde{x}_k} f(\widetilde{r}_k - \widetilde{r}_{k-1} | \widetilde{x}_{k-1}, \widetilde{x}_k).$$

The model is similar if Y is leading and it is observed multiple times between consecutive observations of X.

If $\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1}) \geq L$ and the observation is of the follower, then we expect the location and time of the observation to depend on the last observation of the leader. Towards modeling dependence of the observed locations, we define

$$\delta_k = \frac{\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1}) - \mathrm{dist}(\widetilde{x}_k, \widetilde{y}_k)}{\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1})}. \tag{4}$$

Observe that δ_k, which takes values in the interval $(-\infty, 1]$, measures the relative change in distance between the leader and follower at time t_k. If $\delta_k > 0$ then the follower was observed closer to the leader. We model the distribution over where the follower is observed using δ_k in the following manner. Suppose that, at time t_k, X is leading and an observation is made of Y. Then

$$\mathrm{p}_1(\widetilde{y}_k | \widetilde{x}_{k-1}, \widetilde{y}_{k-1}, m) \propto \begin{cases} 1 + \delta_k & \text{if } \delta_k > -1 \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

where[3] the constant of proportionality is chosen to ensure we have a valid distribution. Note that the transition distribution does not depend on the mixture component m in this case. If $\delta_k \leq -1$ then the distance between the leader and follower has more than doubled since the last observation of the follower. This means that the leader and follower and travelling further apart from each other. In this case, the model above will set the likelihood of a convoy (hypothesis H_1) to zero, and the hypothesis test will declare that the pair of vehicles is not a convoy.

When there is a clear follower (i.e., the distance at time t_{k-1} is greater than L) and the follower is observed, we also expect the inter-observation times of the leader and follower to be dependent. Suppose that X is leading and, at time t_k, we observe Y. Consider the quantity $\widetilde{s}_k - \widetilde{r}_k$ which is strictly positive and gives the time between this observation of Y, the follower, and the last observation of X, the leader. We postulate that this distribution should be such that values of $\widetilde{s}_k - \widetilde{r}_k$ closer to zero are more indicative of the pair being a convoy. A simple way to capture this idea is to model the leader-follower inter-observation time as following the half-normal distribution with parameter $\sigma^2 > 0$,

$$f_{HN}(\widetilde{s}_k - \widetilde{r}_k) = \frac{\sqrt{2}}{\sqrt{\pi\sigma^2}} \exp\left(\frac{-(\widetilde{s}_k - \widetilde{r}_k)^2}{2\sigma^2} \right). \tag{6}$$

To summarize, the forms of the transition distribution (3) for each of the six cases mentioned at the beginning of this subsection are shown in Table I.

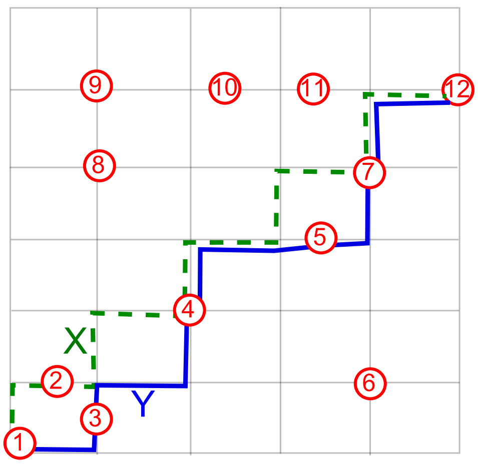
In the convoy model there are still M mixture components in the terms involving the Markov transition matrices $P^{(m)}_{x,x'}$. Therefore the likelihood that a pair of vehicles are traveling as a convoy becomes

$$\mathrm{p}_1(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}) = \max_m \left\{ \mathrm{p}_1(\mathbf{x}_{1:k}, \mathbf{y}_{1:k} | m) \right\}.$$

[3]Note that the above equation is valid even though δ_k in the right-hand side depends on $\widetilde{x}_k$, which appears to be missing from the arguments on the left-hand side. This is because, for the situation considered where the observation at time t_k is of Y, we have $\widetilde{x}_k = \widetilde{x}_{k-1}$, so δ_k is still computable.

TABLE I. VALUE OF THE TRANSITION DISTRIBUTION $p_1(\mathbf{x}_k, \mathbf{y}_k | \mathbf{x}_{k-1}, \mathbf{y}_{k-1}, m)$ FOR THE DIFFERENT CASES CONSIDERED UNDER H_1. HERE $\mathbf{1}\{\cdot\}$ DENOTES THE $0/1$-VALUED INDICATOR FUNCTION, AND THE $\propto$ REFERS TO THE CONSTANT OF PROPORTIONALITY FROM (5).

Vehicle observed at t_k	No Clear Leader ($\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1}) < L$)	Clear Leader ($\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1}) \geq L$) X leading ($\widetilde{r}_{k-1} > \widetilde{s}_{k-1}$)	Clear Leader ($\mathrm{dist}(\widetilde{x}_{k-1}, \widetilde{y}_{k-1}) \geq L$) Y leading ($\widetilde{s}_{k-1} > \widetilde{r}_{k-1}$)
X ($\widetilde{r}_k > \widetilde{s}_k$)	$P^{(m)}_{\widetilde{x}_{k-1}, \widetilde{x}_k} f(\widetilde{r}_k - \widetilde{r}_{k-1} \| \widetilde{x}_{k-1}, \widetilde{x}_k)$	$P^{(m)}_{\widetilde{x}_{k-1}, \widetilde{x}_k} f(\widetilde{r}_k - \widetilde{r}_{k-1} \| \widetilde{x}_{k-1}, \widetilde{x}_k)$	$\propto (1+\delta_k) f_{HN}(\widetilde{s}_k - \widetilde{r}_k) \mathbf{1}\{\delta_k > -1\}$
Y ($\widetilde{s}_k > \widetilde{r}_k$)	$P^{(m)}_{\widetilde{y}_{k-1}, \widetilde{y}_k} f(\widetilde{s}_k - \widetilde{s}_{k-1} \| \widetilde{y}_{k-1}, \widetilde{y}_k)$	$\propto (1+\delta_k) f_{HN}(\widetilde{s}_k - \widetilde{r}_k) \mathbf{1}\{\delta_k > -1\}$	$P^{(m)}_{\widetilde{y}_{k-1}, \widetilde{y}_k} f(\widetilde{s}_k - \widetilde{s}_{k-1} \| \widetilde{y}_{k-1}, \widetilde{y}_k)$



$\mathbf{x_i}$	1		2			4		7		12	
$\mathbf{r_i}$	0		3			8		14		21	
$\mathbf{y_i}$		1		3	4		5		7		15
$\mathbf{s_i}$		1		4	7		10		15		22

Fig. 2. Example of a convoy of two vehicles (X and Y) on a simple grid network. The figure shows the trajectories of each vehicle along the locations of 12 sensors. The table shows the observations (sensor index and observation time) made of both vehicles, spaced so as to help illustrate the sequence of observations over time.

This, as in the independent model, denotes the likelihood of a convoy as the highest likelihood of a convoy for any individual chain.

F. Convoy example

Fig. 2 shows an example convoy scenario where two vehicles, X and Y, transition through a network. The observations of each vehicle are shown in the table. In this example X is leading from times 0 to 4, then Y leads from times 7 to 10, and X leads again from time 14 until the end of the example. The routes taken by the two vehicles are highly correlated but not identical. In addition, the vehicles are not always observed by exactly the same sensors. Thus the example illustrates some of the subtleties we aim to capture in our definition of a convoy.

IV. Convoy Detection via Sequential Hypothesis Testing

Next we discuss our approach to detecting convoys in streams of license plate reads. We consider a typical sequential hypothesis testing setting [8] where the observations arrive successively at the fusion center, ordered by the times r_i and s_i, and after receiving an observation the decision maker must choose from one of three options: 1) declare that the pair of vehicles is a convoy, 2) declare that the pair of vehicles is not a convoy, or 3) wait to receive additional observations. The aim is to make accurate decisions without deferring too long.

For the models described in the previous section, which involve mixtures of Markov chains, to perform testing in a sequential manner we use the sequential generalized likelihood ratio test [18]. The test statistic after $k+1$ total observations is

$$\Lambda(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}) = \frac{\max\limits_{m} \{p_1(\mathbf{x}_{1:k}, \mathbf{y}_{1:k} | m)\}}{\max\limits_{m} \{p_0(\mathbf{x}_{1:k}, \mathbf{y}_{1:k} | m)\}}. \tag{7}$$

The test statistic can be updated in a recursive manner since the individual likelihoods $p_0(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}|m)$ and $p_1(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}|m)$ factorize according to (2). Thus, M likelihood statistics need to be stored and updated for each hypothesis, H_0 and H_1.

Two decision thresholds, η_0 and η_1, are applied so that the decision after each update is given by the well-known rules:

$$\begin{array}{rcccll} & & \Lambda(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}) & < & \eta_0 & \text{decide } H_0 \\ \eta_0 & \leq & \Lambda(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}) & < & \eta_1 & \text{decide “need more data”} \\ \eta_1 & \leq & \Lambda(\mathbf{x}_{1:k}, \mathbf{y}_{1:k}) & & & \text{decide } H_1. \end{array}$$

According to Wald [8], approximate decision regions for the sequential likelihood ratio test can be derived given specific performance criteria: the desired probability of false detection, $P_F \leq \alpha$, and the desired probability of detection, $P_D \geq \beta$, by taking

$$\eta_0 \geq \frac{1-\beta}{1-\alpha} \quad \text{and} \quad \eta_1 \leq \frac{\beta}{\alpha}\,. \tag{8}$$

Using these expressions, with equality, for η_0 and η_1 results in upper and lower bounds on P_D and P_F. This can be used to set the desired performance limitations on the system. Normally in sequential hypothesis testing these bounds will be computed for i.i.d. samples of the two probability densities however the only requirement to achieve these bounds on the sequential test's performance are that the likelihood ratio be able to be decomposed into components which are only dependent on the current sample and the previous likelihood. In a Markov setting the “current” sample is a joint sample of the actual current sample and the previous sample. Therefore since this test can still be decomposed into individual components this analysis still holds.

To evaluate the likelihood models described in this section, parameters of the Markov chain mixture model need to be estimated or configured. These issues are discussed next.

A. Estimation of Markov chain mixture model parameters

In order to use a mixture of discrete Markov chains to more accurately describe the network, the model parameters must be estimated from training data. We use the Expectation Maximization (EM) algorithm [19] for this purpose. Previous work for estimation of a mixture of Markov chains using EM addressed the problem in the setting where each observation is an individual transition that may come from a different mixture component [20]. For the observations considered here, we assume that each vehicle's entire trajectory is associated with a single (latent) mixture component (rather than each observed transition of each vehicle potentially coming from a different mixture component). The number of mixture components can be determined using standard measures for goodness of fit in model order selection, such as the *Bayes Information Criterion* (BIC) [21].

B. Comments on the leader-follower inter-observation time distribution under H_1

The parameter σ^2 of the half-normal distribution appearing in (6), used in the likelihood model under H_1, also needs to be specified. To consider a pair of vehicles to be driving as a convoy, one would like that the vehicles do not drift too far away from each other. We take $\sigma^2 = 30$ in the experiments, roughly corresponding to a maximum allowable time separation of 100 seconds between observations the leader and follower under H_1. To see this correspondence, note that integrating the half-normal pdf from 0 to 100 is close to 1 when $\sigma^2 = 30$.

C. Other system parameters

For practical reasons, tracks of pairs of vehicles are only started when two vehicles are first seen close together in distance ($< L$) and in time. This threshold, which is also used in the statistical test, controls how far apart vehicles can drive in parallel routes while still being considered a convoy. It also controls how close together vehicles need to get in order to start the statistical test. We introduce two additional time threshold parameters, T_s and T_d. The parameter T_s is used to determine when to begin tracking a given pair of convoy vehicles (i.e., running the sequential hypothesis test for the given pair). A test is started if the vehicles are observed at locations at most a distance of L apart within T_s time units. The choice of T_s will only control when tests start. A logical choice for this parameter might be related to the choice of the 95% confidence interval of the half-normal distribution. For example, $\sigma^2 = 30$ results in approximately 100 as the maximum value for the pdf of the half-normal in the 95% area. Therefore a logical choice to mimic the convoy sequential test might be 100 seconds. Setting this value very large would trigger the start of a lot of unnecessary tests, tracking pairs of vehicles, which would likely terminate after a few observations are made. The parameter T_d is introduced for practical reasons, to also limit the number of consecutive sequential likelihood ratio tests being evaluated; if T_d time units have elapsed and no new observation of either of the vehicles considered in a test has been received, then that track is terminated. This is the same as the track of the vehicles getting lost since they likely have travelled outside the field of view of the sensor network or at least one vehicle has parked and therefore will not be observed by the network.

V. Experimental Evaluation

A. Data description

Next we study the performance of the proposed sequential hypothesis test using the models described in Section IV against simulated data. A regional traffic assignment model for the Montreal metropolitan area is described in Sider et al. [22]. The model takes as an input the 2008 Origin-Destination (OD) trip data for the Montreal region provided by Montreal's *Agence Métropolitaine de Transport* and assigns it on the network using a stochastic assignment in the VISUM platform [23]. The regional network consists of 127,217 road links and 90,467 nodes associated with over 1500 traffic analysis zones. It also contains various road characteristics such as the type, length, speed limit, capacity, and number of lanes [22]. Note that this model has been validated using both traffic counts [24] and speed data collected using GPS [25].

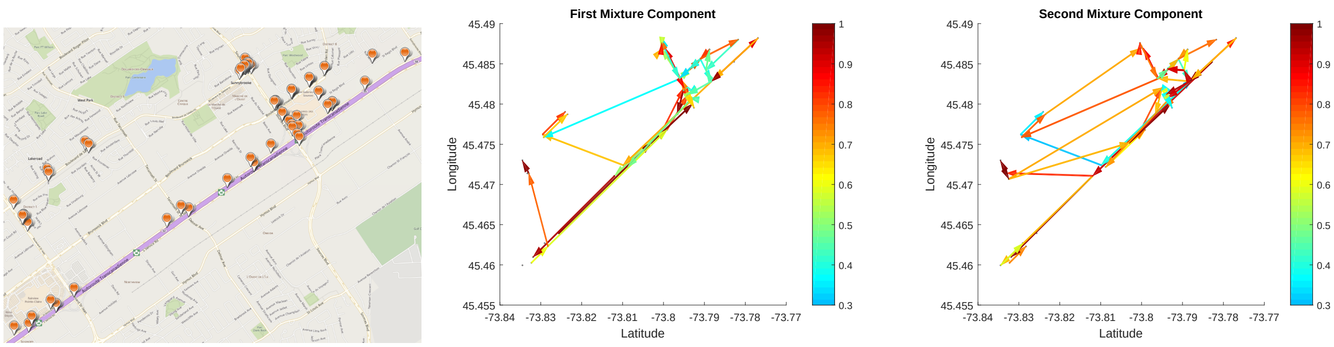
Output from the traffic assignment simulations consists of an array that contains a detailed description of all paths connecting pairs of origin-destination zones for every hour of the day. Using this load information, we simulate a population of 2 million vehicles (roughly the number of registered vehicles in the greater Montreal region). These vehicles are sent randomly from zone to zone at random times during each hour along the paths from the Sider et al. [22] dataset, with the number of vehicles per path chosen to match the prescribed loads.

Sensors are placed at the 75 locations shown in Fig. 3(a). Each sensor records the identification number (license plate) of the vehicles as they pass by the sensors' locations. The data recorded by these sensors constitutes the baseline, normal traffic used in our experiments.

Two datasets were then simulated on this sensor network. Each simulation results in 24 hours of data and contains approximately 500,000 observed vehicles. The first of these two simulations was used for training, to fit the parameters of the mixture of Markov chains as well as the parameters to the distribution describing the time transitions. The second dataset was then used as a test dataset in which convoys of varying types were injected along with vehicles traveling independently. Performing a cursory analysis on each dataset we note that each vehicle is observed nine times, on average. This means that any detections which will occur only have access to a limited amount of data from each vehicle in the timespan the vehicle is present in the data. This dataset is the basis for the performance analysis reported later this section.

B. Estimated Transition Matrices

To fit the transition model parameters used in the simulations, multiple iterations of the EM algorithm were run while varying the number of mixture components in order to estimate the Markov transition matrices and initial distributions. The Bayesian Information Criteria (BIC) [21] was used for model order selection. More specifically, for each possible number of mixture components in the range $\{1, 2, \ldots, 5\}$, the EM algorithm was executed from fifty different random initializations.



(a) (b) (c)

Fig. 3. (a) Locations of the 75 simulated sensors along a stretch of Highway 40 in Montreal, Canada. Along this stretch there are two exits, one near the top-right corner (where sensors are shown perpendicular to the highway) and the other near the bottom-left corner of the figure. Sensors are also located on the feeder roads that run along side of the highway. The image appears to have less than 75 sensors since many of the points represent 2 sensors (one pointed in each direction). This is necessary since LPR sensors require that they be monitoring a specified direction so to cover a bi-directional road two sensors are necessary. A Markov chain mixture model is fit to simulated traffic from a 24-hour training period, and it is determined that a two-component mixture model provides the best fit as measured using the BIC. The transition matrices of these two mixture components are shown in panels (b) and (c). It can be seen that, while mixture components capture the flow of traffic along the highway, they capture distinctly different trends in terms of traffic entering/exiting the highway, and off of the highway.

We did not try more than 5 mixture components since we noted after multiple trials that the BIC for mixtures with more components got worse, rapidly. For this network, the model with the best BIC across all 50×5 random initializations is a mixture with 2 components. The two estimated transition matrices are visualized in Figs. 3(b) and 3(c). Each of the estimated components exhibits essentially the same behavior on the highway between exits. This is reasonable, since a vehicle traveling down the highway without a possible exit will continue traveling in the same direction. The differences in the transition matrices can be more aptly visualized on the side-roads off the highway. We can see that different traffic patterns are captured in these small offshoots from the highway.

C. Inter-observation time distribution under H_0

In addition to the Markov chain mixture model, the distribution governing the inter-observation times needs to be specified. As mentioned in Section III-A, a valid distribution for inter-observation times should have support on $\mathbb{R}_+$. For this work we estimate a time transition based on the starting state using various exponential family models. Using the dataset described above, the normal distribution, inverse-Gaussian, and gamma distributions were fit to the data. Using the BIC as a measure of goodness, the heavy-tailed nature of the inverse-Gaussian distribution provided the best fit to the training data. Thus, we take $f(\tau|x, x')$, the likelihood that the time between two consecutive observations of a vehicle is τ time units given it was observed at sensor x and then at sensor x' (after τ time units), to be the inverse-Gaussian distribution,

$$\begin{aligned} & f_{IG}(\tau; \mu_{x,x'}, \lambda_x) \\ & \quad = \left[\frac{\lambda_x}{2\pi\tau^3}\right]^{1/2} \exp\left[\frac{-\lambda_x(\tau - \mu_{x,x'})^2}{2\mu_{x,x'}^2 \tau}\right] \mathbf{1}\{\tau \geq 0\}, \end{aligned}$$

where $\mathbf{1}\{\cdot\}$ is the 0/1-valued indicator function, $\mu_{x,x'}$ is the mean time to transition from state x to state x', and λ_x is the shape parameter associated with trajectories departing state x. When viewed as a generalized linear model [26], the inverse-Gaussian distribution has link function

$$\begin{aligned} \frac{1}{\mu_{x,x'}^2} &= \alpha_x + \mathrm{dist}(x, x')\beta_x \\ \mu_{x,x'} &= \frac{1}{\sqrt{\alpha_x + \mathrm{dist}(x, x')\beta_x}} \end{aligned}$$

where, now, α_x, β_x, and λ_x are the parameters to be estimated, and $\mathrm{dist}(x, x')$ is the distance between states x and x'. These parameters are estimated from the training data using Fisher scoring [26].

D. Simulating convoys

The simulated dataset described in Section V-A is intended to represent normal background traffic. While we cannot guarantee there are no instances of convoys in this dataset, the appearance of any is unintentional. In order to evaluate the performance of the proposed sequential hypothesis testing approach, we inject convoys into the background data. Simulation of convoys involves determining two main factors: 1) the trajectories that will be taken by the vehicles, and 2) how the spacing between them will evolve over time. We consider two possibilities for each of these factors.

For the trajectories, in one case we simulate a convoy where the leader remains fixed for the entire trajectory and the follower takes exactly the same trajectory as the leader, where the leader's trajectory is sampled from one of the Markov chain mixture components. Alternatively, to allow for the leader and follower to take slightly different paths, we also simulate convoys where the follower's trajectory is sampled using the model described in Section III-E, e.g., using (5).

To determine the timing between when the leader and follower are observed, we also consider two possibilities. In one case, the follower is always observed exactly one second after the leader. At a typical highway speed of 100 km/h, separation of 1 second corresponds to a distance of 27.8 meters between the vehicles, or 5–6 car lengths. Alternatively, we also simulate convoys where the follower's observation times are sampled from the half-normal distribution with parameter $\sigma^2 = 30$, following the model proposed in Section III-E. The value $\sigma^2 = 30$ was chosen to allow an approximate maximum of 100 seconds of separation between vehicles in a convoy. If one solves the equation

$$1 = \int_0^{100} f_{HN}(y|\sigma^2)dy \tag{9}$$

for σ^2, one gets a value of approximately $\sigma^2 \approx 30$. This is a parameter to be chosen which allows for a target allowed maximum separation time between vehicles which the detection method will be sensitive to.

Taking all possible combinations of the two trajectory models and timing models described above leads to four ways in which convoys may be simulated. These four scenarios are summarized in Table II, and all four are considered in the simulation results discussed below. Convoys of the varying types are simulated for approximately 18 observations (9 of each vehicle) and last anywhere from a few seconds to about 30 minutes, depending on the road segment they were randomly started on. Although we simulate such longer-lasting convoys, in Section V-H we study the average number of observations required by the sequential hypothesis testing procedure to make a decision to better understand how many observations are required and how this number depends on the performance criteria P_F and P_D.

To simplify the presentation, for the rest of this section we only present and discuss results for convoys simulated according to Scenario 4. Results for the other three scenarios, which are included in the appendix, are qualitatively very similar.

E. Probability of detection

To assess the probability of detection of the proposed sequential test, we simulate 1000 convoys for each of the four scenarios described in Table II, and we evaluate empirical probability of detection as a function of the decision thresholds η_0 and η_1. Fig. 4 shows the probability of detection at the time of the first decision for Scenario 4. Varying the threshold η_0 has relatively little effect, especially for $\ln(\eta_0) < -5$. Setting the thresholds according to (8) with design criteria $\alpha = 0.0111$ and $\beta = 0.9999$ gives $\ln(\eta_0) \approx -9.20$ and $\ln(\eta_1) \approx 4.50$, for which the resulting probability of detection is $P_D = 0.9332$.

F. Probability of false detection

We next simulate 1000 pairs of vehicles traveling through the road network independently. Each pair is simulated according to the same mixture component in the mixture of Markov chains and are sampled, as with the convoy case, for 18 observations (9 of each vehicle). The spacing of these

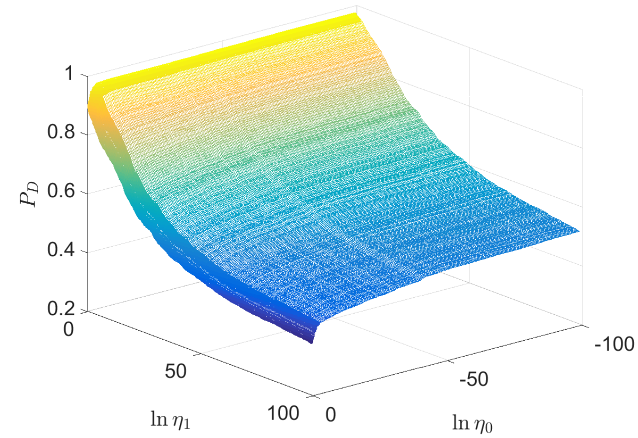


Fig. 4. Probability of detection for varying decision boundaries η_0 and η_1 with convoys simulated by Scenario 4.

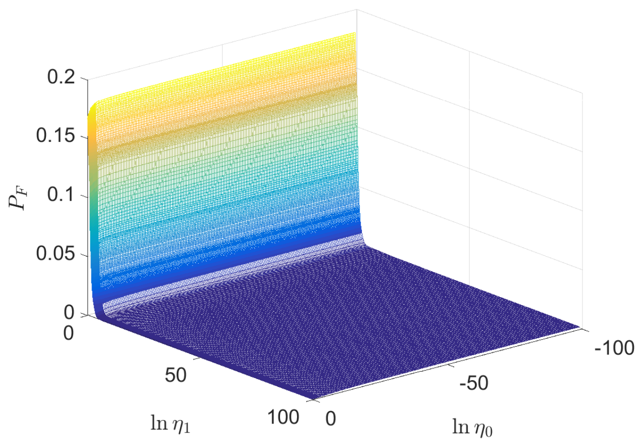


Fig. 5. Probability of false detection for varying decision boundaries η_0 and η_1 for vehicles simulated following the independent model

observations in time depends on the random starting location and the network links traveled. We use these to study the probability of false detection for different values of $\ln \eta_0$ and $\ln \eta_1$. Fig. 5 shows the empirical probability of false detection as $\ln(\eta_0)$ and $\ln(\eta_1)$ are varied. As can be seen, the probability of false detection quickly drops to an almost negligible amount with a small increase in $\ln \eta_1$. Using the same decision bounds mentioned above, $\ln(\eta_0) = -9.20$ and $\ln(\eta_1) = 4.50$, the probability of false detection is $P_F = 0.0031$.

Fig. 6 shows a scatter plot of P_D versus P_F, where each filled point corresponds to a particular choice of η_0 and η_1. The color of each point corresponds to the value of η_0. As is evident from the plot, as η_0 tends to $-\infty$, the probability of detection increases. One can also see subsets of points falling in roughly vertical groups. These correspond to the performance of the test when η_1 is held fixed and η_0 is varied, giving similar values of P_F while varying P_D.

TABLE II. SIMULATED CONVOY CONFIGURATIONS

	Time separation between X and Y	Discrete Transition Model
Scenario 1	Constant separation of 1 second	X strictly followed by Y
Scenario 2	Constant separation of 1 second	X and Y following model in Section III-E
Scenario 3	$\sim$ HalfNormal($\sigma^2 = 30s$)	X strictly followed by Y
Scenario 4	$\sim$ HalfNormal($\sigma^2 = 30s$)	X and Y following model in Section III-E

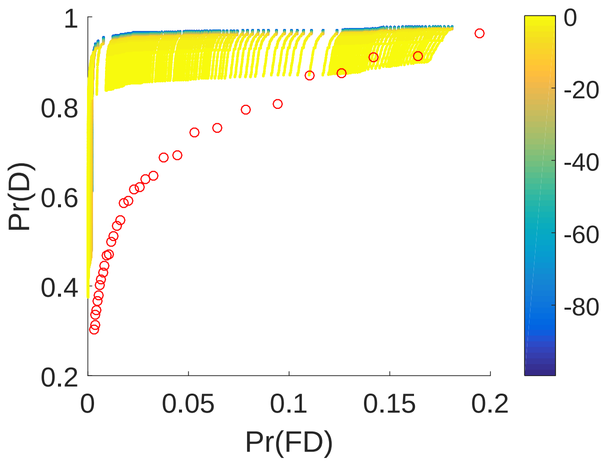


Fig. 6. Scatter plot of the resulting probability of false detection values versus the probability of detection values for all combinations of $\ln(\eta_0)$ and $\ln(\eta_1)$ where convoys are simulated using the convoy model described in Section III-E. The vertical coloring denotes changes in η_0. Note that the horizontal axis (P_F) ranges from 0 to 0.2. Overlayed in red are the probability of detection and false detection rates from the thresholding approach.

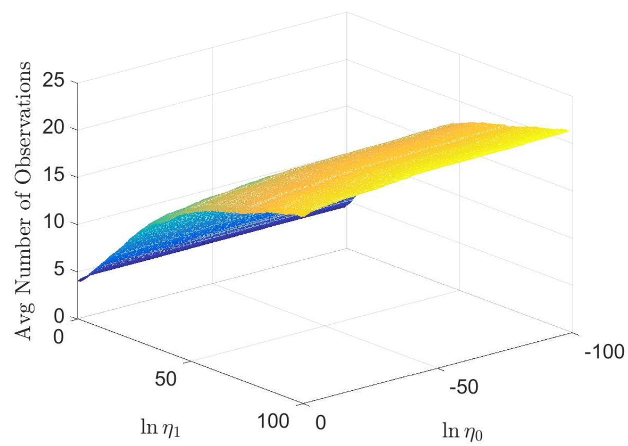


Fig. 7. Expected number of observations to make a decision under H_1 for varying decision boundaries η_0 and η_1 with convoys simulated by Scenario 4.

G. Comparison to a Simple Thresholding Approach

We compare the proposed method with a simple thresholding approach. A threshold is directly applied to the total number $n(t)$ of observations of a pair of vehicles, based on the intuition that the more often a pair of vehicles are observed near each other, the more likely they are to be a convoy. For a fair comparison, we apply the same system parameters as described in Section IV: to first consider a pair of vehicles as a potential convoy they need to be observed within a distance of L from each other within T_s time units, and to continue being considered as a potential convoy the pair must be observed very subsequent T_d time units afterwards.

The empirical detection probability and false alarm probability of the thresholding approach are also shown in Fig. 6 as red hollow circles. The threshold on $n(t)$ is varied from 2 to 40. (Note that $n(t)$ only takes values in the positive integers, so we only apply integer thresholds.) When a small threshold is used, the simple thresholding approach achieves a P_D comparable to what can be achieved using the proposed approach, but with a very high probability of false detection (nearly 0.2). Increasing the threshold reduces both the probability of false detection and the probability of detection. In general, for very low probability of false detection, which is clearly desirable in applications, the proposed approach has a significantly higher P_D. Moreover, it is evident from Fig. 6 that the performance of the proposed approach is much less sensitive to the choice of threshold parameters η_0 and η_1.

H. Expected number of observations to make a decision

In addition to making accurate decisions (low P_F and high P_D), it is important to understand how varying the decision thresholds of the sequential hypothesis test affects the number of observations required to make a decision. Figs. 7 and 8 show the average number of observations ($n(t)$, the total number of observations of either vehicle) to make a decision under H_1 and H_0, respectively, as a function of the decision thresholds. A smaller value in this metric is better since it corresponds to a faster time to detect convoys under H_1, and a faster time to stop tracking non-convoy pairs under H_0. In a practical implementation, discarding non-convoy pairs quickly (without sacrificing accuracy in terms of P_D and P_F) is desirable since the computational resources used by the sequential hypothesis test (both memory and CPU cycles) are proportional to the number of pairs of vehicles being tracked.

As $\ln(\eta_0) \to -\infty$ and $\ln(\eta_1) \to \infty$, the number of observations required to make a decision for H_1 increases. Focusing on the specific decision threshold values $\ln(\eta_0) = -9.20$ and $\ln(\eta_1) = 4.50$ mentioned before, Figs. 9 and 10 show histograms of the number of observations required to make a decision under H_1 and H_0, respectively. In both cases, decisions are made, on average, when roughly 10–12 total observations of the pair of vehicles are available (i.e., 5–6 observations of each vehicle).

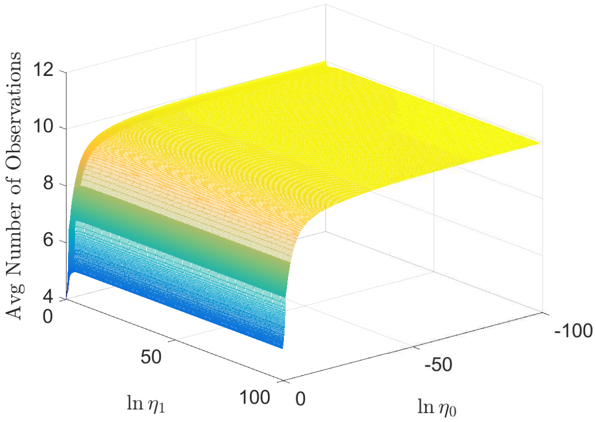


Fig. 8. Expected number of observations to make a decision under H_0 for varying decision boundaries η_0 and η_1 where vehicles are simulated independent of each other.

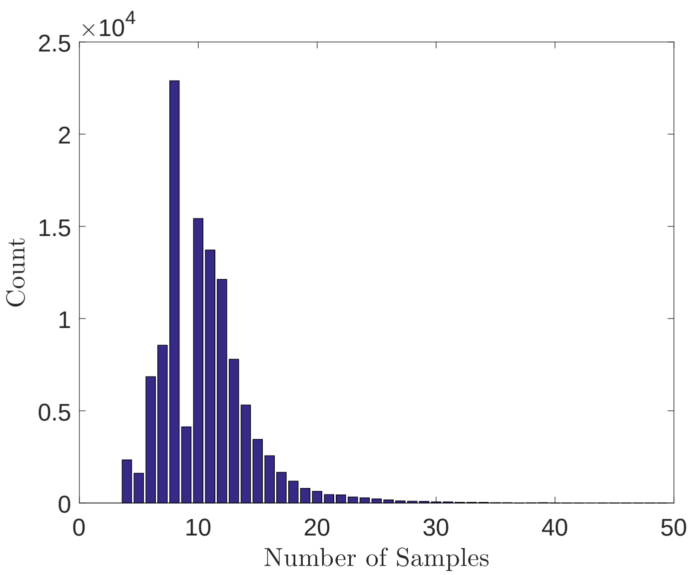


Fig. 9. Histogram of the number of samples to make a decision under the alternate hypothesis (H_1) where convoys were simulated with half-normally distributed time separation and following the discrete convoy model in Section III-E.

VI. Discussion

A. Regarding the explicit use of road network data

The sequential detection approach adopted in this paper does not explicitly make use of knowledge of the road network topology. Instead, it is implicitly encoded in the transition matrices of the Markov chain mixture model. Such information could be used in the models, e.g., when calculating the distance $\mathrm{dist}(x, x')$, if it is available. Tracking vehicles explicitly over a state space consisting of the entire road network would be computationally cumbersome in a large system (which may observe on the order of tens of thousands of vehicles per hour), and a system making use of detailed road maps would also

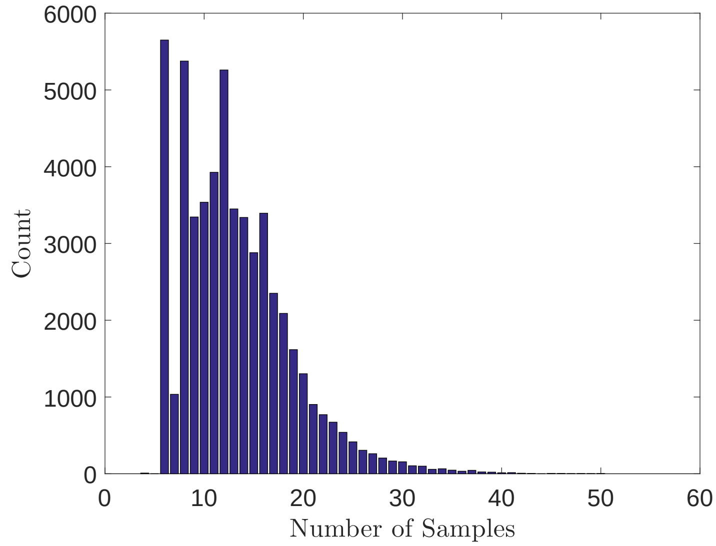


Fig. 10. Histogram of the number of samples to make a decision under the independent hypothesis (H_0) where vehicles are traveling independently.

require updating of the maps when segments are closed (e.g., for construction) or changed (e.g., re-zoning). On the other hand, the proposed approach implicitly models traffic patterns using the Markov chain mixture model. The parameters of this model can be estimated directly from the data, and so no additional input or tuning is required.

B. Detecting convoys of more than two vehicles

In order to detect if a group of vehicles larger than two are traveling as a convoy a simple post-analysis can be performed. In order to understand this post-analysis for groups of convoys, consider a target vehicle X and suppose that we detect N other vehicles as traveling in a convoy with X at a specific time. We then simply look at these N vehicles which were detected as in a convoy with X and look if they were also detected as being in a convoy with each other. This creates a set of vehicles where all the pairwise combinations are detected to be in a convoy in a set timeframe. This is a detected convoy "group". We note that the approach just described can be related to the notion of density-connected sets used in [13].

One may be tempted to view the problem of detecting convoys of more than two vehicles as a sort-of graph partitioning or community detection problem, with vertices in the graph corresponding to vehicles and edges placed between two nodes that belong to a convoy. The pairwise test presented in this paper identifies where there are likely edges, and one would hope that a convoy of two or more vehicles would give rise to dense connections between the vehicles in the convoy. However this is not necessarily the case since convoys may be formed by long lines of vehicles (e.g., along a single-lane road). For example, if three vehicles, $X - Y - Z$, form a convoy our test may not detect the correlation between X and Z directly if they are too far apart. This presents one of the main challenges we anticipate with detecting convoys of more than two vehicles. We leave a more detailed study and in-depth analysis of detecting larger convoy groups to future work.

C. Using different estimated network properties for different times of the day

Some other issues which might arise in practice are such things as accidents or road closures as well as how traffic patterns behave differently throughout the day (e.g. rush hour). All of these real-world issues will cause traffic to behave differently from the network which was originally trained on. The problem of random events such as accidents and road closures is difficult to handle due to the unpredictable nature of it. This will likely cause more anomalies (such as convoys) to be flagged in the algorithm due to more vehicles taking a lower-likelihood route.

However the case of a varying traffic pattern throughout the day is one which is much more simple to mitigate. By simply swapping out the transition matrices as well as the properties for the inverse-Gaussian distributions and the initial distributions, one can in real-time update the detector for more realistic traffic patterns. This could be done, say, every hour to mimic changing traffic patterns throughout the day. This would not change the algorithm's design since it would say simply for a specific tracked pair of vehicles "the first n samples came from the 1 a.m. to 2 a.m. mixture while the next m samples came from the 2 a.m. to 3 a.m. mixture". This allows the algorithm to handle even a continuous-time distribution for the underlying mixture of Markov chains. An analytic solution will become much more difficult due to the addition of many additional chains to estimate (possibly infinite in the continuous-time mixture case), however it might be able to drastically improve the detection and false detection performance by more accurately measuring the nominal traffic distribution.

VII. Conclusion

This paper proposes a novel approach to detecting convoys in urban environments. Typically long-range sensors are not applicable in urban environments. This means that only by using short-range sensors such as LPR can one do many types of road network analysis including convoy detection. The algorithm presented only uses a small amount of information about the detected vehicles to perform convoy detection which is an added benefit for minimizing the computational complexity. It is also capable of detecting convoys in real time as data arrives.

In the problem formulation of this work we assumed that measurements are exact; there are no mis-read license plates and no missed license plate reads. Our future work will address the case of missing and noisy data using a hierarchical Bayesian approach by adding one layer, so that the vehicle trajectory model becomes a hidden mixture of Markov chains.

This paper focused on detecting convoys of vehicles in a road network. Individual vehicles were modeled as moving along paths in the network according to a first-order Markov model, and convoys are two or more vehicles whose paths are correlated in space and time. An interesting extension of this approach would be to detect when two or more epidemics spreading over a network are correlated. First-order Markov models are also commonly used to model epidemics spreading over networks, but the resulting patterns are trees rather than paths. In future work it would be interesting to explore extensions of the sequential hypothesis testing framework considered in this paper for detecting correlated epidemics.

Appendix

Fig. 11(a) shows P_D as a function of the decision thresholds when convoys are simulated using Scenario 1. This situation is where a vehicle X moves independently through the network while vehicle Y follows exactly the same path as X with a 1-second lag. It can be seen here that the detection accuracy degrades quickly with the increase of the $\ln \eta_1$. This appears to no longer be the case in the next scenario, Scenario 2, as shown in Fig. 11(b) where there is still a constant 1-second time separation but the transitions of Y are following the convoy model from Section III-E. This is because in scenario following the model from Section III-E one vehicle can, in many circumstances, take an alternate, lower likelihood, path which is close to the leader so the likelihood of H_0 drops faster than the likelihood of H_1. For example, consider two vehicles traveling on parallel paths where one vehicle is on a high-likelihood path (such as a highway) and another is on a lower-likelihood path (such as a service road parallel to the highway). In this case the vehicle on the lower likelihood path will make the likelihood of H_0 lower faster than the likelihood of H_1 decreases.

One can also see that the mitigation of the fast drop in the shape of the surface in Fig. 11(a) can be likely attributed to the constant time separation of 1 second as in Figs. 11(c) and 4. Here the exponential drop in the probability of detection with the increase of $\ln(\eta_1)$ appears to become at worst a linear relationship. This means that allowing a floating leader along with a variable distance between vehicles increases our detection ability. This is very good news since a constant separation between vehicles of 1 second throughout an entire observation sequence is very unlikely.

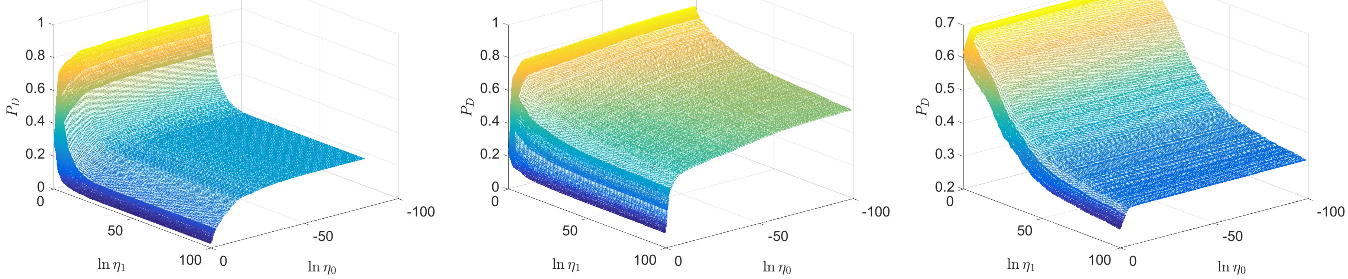
Figs. 12(a), 12(b), and 12(c) show the average number of observations required to make a decision under H_1 when convoys are simulated according to Scenario 1, 2, and 3, respectively. These figures exhibit a similar trend to that presented in Fig. 7. The main difference is in terms of the rate at which the average number of decisions plateaus with changes of $\ln \eta_1$.

References

[1] S. Lawlor and M. Rabbat, "Detecting convoys in networks of short-range sensors," in *Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 2014.

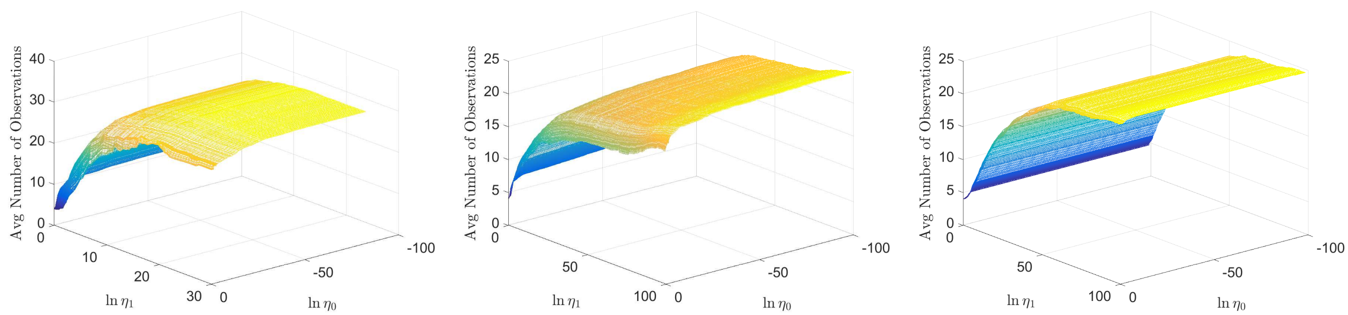
[2] A. Homayounfar, A. Ho, N. Zhu, G. Head, and P. Palmer, "Multi-vehicle convoy analysis based on ANPR data," in *Intl. Conf. on Imaging for Crime Detection and Prevention*, Nov. 2011, pp. 1–5.

[3] S. van de Hoef, K. Johansson, and D. Dimarogonas, "Fuel-optimal centralized coordination of truck-platooning based on shortest paths," in *American Control Conf.*, Chicago, IL, Jul. 2015.



(a) (b) (c)

Fig. 11. Probability of detection for varying decision boundaries η_0 and η_1 with convoys simulated by (a) Scenario 1, (b) Scenario 2, and (c) Scenario 3.



(a) (b) (c)

Fig. 12. Expected number of observations to make a decision under H_1 for varying decision boundaries η_0 and η_1 with convoys simulated by (a) Scenario 1, (b) Scenario 2, and (c) Scenario 3.

[4] A. Alam, B. Besselink, V. Turri, J. Mårtensson, and K. Johansson, "Heavy-duty vehicle platooning for sustainable freight transportation: A cooperative method to enhance safety and efficiency," *IEEE Cont. Sys. Mag.*, vol. 35, no. 6, pp. 34–56, Dec. 2015.

[5] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Trans. on Sig. Proc.*, vol. 51, no. 8, pp. 2191–2204, Aug. 2003.

[6] A. Hero, "Geometric entropy minimization (GEM) for anomaly detection and localization," in *Conf. on Neural Information Processing Systems*, Vancouver, Canada, Dec. 2010.

[7] C. Scott and E. Kolaczyk, "Nonparametric assessment of contamination in multivariate data using generalized quantile sets and FDR," *J. of Computational and Graphical Stat.*, vol. 19, no. 2, pp. 439–456, Jun. 2010.

[8] A. Wald, *Sequential Analysis*, ser. Wiley Publication In Statistics, R. A. Bradley, J. S. Hunter, D. G. Kendall, and G. S. Watson, Eds. John Wiley & Sons, Inc., 1966.

[9] W. Koch, "Information fusion aspects related to GTMI convoy tracking," in *Fifth Int. Conf. on Information Fusion*, vol. 2, 2002, pp. 1038–1045.

[10] E. Pollard, B. Pannetier, and M. Rombaut, "Convoy detection processing by using the hybrid algorithm (GMCPHD/VS-IMMC-MHT) and dynamic Bayesian networks," in *Int. Conf. on Information Fusion*, vol. 12, Seattle, WA, USA, July 2009.

[11] E. Pollard, M. Rombaut, and B. Pannetier, "Bayesian networks vs. evidential networks: An application to convoy detection," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Dortmund, Germany, Jun. 2010.

[12] C. S. Jensen, D. Lin, and B. C. Ooi, "Continuous clustering of moving objects," *IEEE Trans. on Knowledge and Data Eng.*, vol. 19, no. 9, pp. 1161–1174, Sept 2007.

[13] H. Jeung, M. L. Yiu, X. Zhou, C. S. Jensen, and H. T. Shen, "Discovery of convoys in trajectory databases," in *Intl. Conf. on Very Large Data Bases*, Auckland, New Zeland, Aug. 2008, pp. 1068–1080.

[14] P. Kalnis, N. Mamoulis, and S. Bakiras, "On discovering moving clusters in spatio-temporal data," in *Intl. Symp. on Spatial and Temporal Databases*, Angra dos Reis, Brazil, Aug. 2005, pp. 364–381.

[15] T. Weiherer, E. Bouzouraa, and U. Hofmann, "A generic map based environment representation for driver assistance systems applied to detect convoy tracks," in *IEEE Intl. Conf. on Intelligent Transportation Systems*, Anchorage, AK, Sep. 2012.

[16] J. Yeoman and M. Duckham, "Decentralized network neighborhood information collation and distribution for convoy detection," in *Seventh Int. Conf. on Geographic Information Science*, Columbus, OH, September 2012.

[17] R. A. Howard, *Dynamic Probabilistic Systems : Semi-Markov and Decision Processes*, 1st ed. Dover Publications, 2007.

[18] M.-C. Shih, T. L. Lai, J. F. Heyse, and J. Chen, "Sequential generalized likelihood ratio tests for vaccine safety evaluation," *Stat. in Medicine*, vol. 29, no. 26, pp. 2698–2708, November 2010.

[19] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statatistics Society Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[20] T. J. Perkins, "Maximum likelihood trajectories for continuous-time Markov chains," in *Advances in Neural Information Processing Systems 22*, 2009, pp. 1437–1445.

[21] S. T. Buckland, K. P. Burnham, and N. H. Augustin, "Model selection:

An integral part of inference," *Biometrics*, vol. 53, no. 2, pp. 603–618, 1997.

[22] T. Sider, A. Alam, M. Zukari, H. Dugum, N. Goldstein, N. Eluru, and M. Hatzopoulou, "Land-use and socio-economics as determinants of traffic emissions and individual exposure to air pollution," *J. of Transport Geography*, vol. 33, no. 0, pp. 230 – 239, 2013.

[23] PTV Vision, *PTV Vision, 2009*, VISUM 11.0 Basics ed., PTV AG, Karlsruhe, Germany, 2009.

[24] T. Sider, A. Alam, W. Farrell, M. Hatzopoulou, and N. Eluru, "Evaluating vehicular emissions with an integrated mesoscopic and microscopic traffic simulation," *Canadian J. of Civil Eng.*, vol. 41, no. 10, pp. 856–868, Aug. 2014.

[25] A. Alam, G. Ghafghazi, and M. Hatzopoulou, "Traffic emissions and air quality near roads in dense urban neighborhoods: Using microscopic simulation for evaluating effects of vehicle fleet, travel demand, and road network changes," *J. of the Transportation Research Board*, vol. 2427, pp. 83–92, 2014.

[26] A. Agresti, *Categorical Data Analysis*, ser. Wiley Series in Probability and Statistics. Wiley, 2013.