Understanding City Traffic Dynamics Utilizing Sensor and Textual Observations

Pramod Anantharam, Krishnaprasad Thirunarayan, Surendra Marupudi, Amit Sheth, Tanvi Banerjee
Kno.e.sis, Wright State University, Dayton OH, USA
{pramod,tkprasad,surendra,amit,tanvi}@knoesis.org

Abstract

Understanding speed and travel-time dynamics in response to various city related events is an important and challenging problem. Sensor data (numerical) containing average speed of vehicles passing through a road link can be interpreted in terms of traffic related incident reports from city authorities and social media data (textual), providing a complementary understanding of traffic dynamics. State-of-the-art research is focused on either analyzing sensor observations or citizen observations; we seek to exploit both in a synergistic manner.

We demonstrate the role of domain knowledge in capturing the non-linearity of speed and travel-time dynamics by segmenting speed and travel-time observations into simpler components amenable to description using linear models such as Linear Dynamical System (LDS). Specifically, we propose Restricted Switching Linear Dynamical System (RSLDS) to model normal speed and travel time dynamics and thereby characterize anomalous dynamics. We utilize the city traffic events extracted from text to explain anomalous dynamics. We present a large scale evaluation of the proposed approach on a real-world traffic and twitter dataset collected over a year with promising results.

Introduction

There is an increasing body of research on understanding traffic flow for efficient management of mobility in a city. Currently, there are over 1 billion cars on roads and this number is expected to double by 2020 (IBM 2014). Vehicular traffic increased by 236% from 1981 to 2001 while the world population grew only by 20% (IBM 2014). Increased urbanization has impacted the mobility of people in cities. Zero traffic fatalities and minimizing traffic delays are some of the grand challenges in Cyber-Physical Systems (Rajkumar et al. 2010). To overcome these challenges, we need a deeper understanding of the interactions between various events in a city and its impact on traffic. Current traffic assessment techniques focus on analysis of fine grained sensor observations to predict delays (Ko and Guensler 2005; Anderson and Bell 1997; De Fabritiis, Ragona, and Valenti 2008; Sun, Zhang, and Yu 2006). However, these are limited and opaque, and do not explain the reasons for traffic flow variations. Increasingly, observations of real-world systems span Physical, Cyber and Social (PCS) domains with heterogeneous and multi-modal observations (Sheth, Anantharam, and Henson 2013). For example, observations related to traffic spans both sensor and textual modality. We believe that social media data (Goodchild 2007; Nagarajan, Sheth, and Velmurugan 2011) can provide better understanding of speed dynamics by providing information complementary to sensor data. While social data has played a major role in interpreting situations and applications such as political debates, civil unrest, crime prediction, disaster relief and coordination (Crooks et al. 2013; Boulos et al. 2011; Sakaki, Okazaki, and Matsuo 2010), it has been under-utilized in understanding PCS systems. Extracting traffic related information from social data has been carried out by some studies (Wanichayapong et al. 2011; Zeng et al. 2013).

Important challenges in modeling and explaining speed and travel time variations, collectively called traffic dynamics, include: i) *Modeling nonlinear traffic dynamics* due to various city events, temporal landmarks such as peak-hour and off-peak hour traffic, and random noise that influence traffic, ii) *Integration of heterogeneous data sources* spanning both social and sensor data, iii) *Scalability issue* involving large data size due to continuous data collection from sensors, iv) *Identifiability of traffic events* due to limited modality of traffic related observations. E.g., many events effect the average speed of vehicles passing through a road link, so: can we even identify various events by just studying traffic dynamics? and v) *Uncertain impact* due to contextual dependency. E.g., events may manifest in speed variations only when a road link has high enough traffic volume.

We address the challenging task of modeling non-linear traffic dynamics utilizing a Restricted Switching Linear Dynamical System (RSLDS) for learning normalcy models. RSLDS captures the non-linearity of speed and travel-time dynamics by segmenting speed and travel-time observations into simpler components amenable to description using linear models such as Linear Dynamical System (LDS) (Bishop 2006). We provide a rationale for using LDS to model traffic dynamics by motivating the modeling needs and connecting it to the capabilities of LDS. We address the heterogeneity challenge by using events extracted from tweets and available formal traffic reports (textual data) to explain anomalies in traffic dynamics (sensor data). We deal with the large data size challenge through a scalable implementation of our approach on Apache Spark.

Our work is one of the first efforts in associating textual observations with sensor anomalies in the domain of traffic. Using our approach, we address research questions such as: Do the traffic events reported by city authorities manifest in the speed and travel time variations? How do we formally model speed and travel time dynamics? How do we capture normalcy and thereby characterize anomalies? Can we utilize city events extracted from tweets to explain anomalies?

Related Work

Related work can be organized into generic and time series based approaches based on the modeling principles employed.

Generic Approaches

Sustainability researchers are studying traffic conditions using sensors on road and GPS sensors on vehicles to predict congestion. Current research on traffic data analytics predominantly uses a single modality such as sensor data for understanding delays (Ko and Guensler 2005; Lee, Tseng, and Tsai 2009; Anderson and Bell 1997; Pattara-Atikom, Pongpaibool, and Thajchayapong 2006; De Fabritiis, Ragona, and Valenti 2008; Sun, Zhang, and Yu 2006). Work on traffic diagnostics connects events to congestions utilizing historical data and applies it to the near real-time observations for explaining congestions in terms of city events (Lécué, Schumann, and Sbodio 2012; Daly, Lecue, and Bicer 2013). Inferring the root cause of traffic congestion is investigated by (Chawla, Zheng, and Hu 2012). The origin and destination of a car is modeled as a latent variable and the flow of cars observed from GPS data is modeled as the observed variable. The root cause does not include the city events that may influence traffic and even cause a change in the origin and the destination of cars. (Horvitz et al. 2012) use a Bayesian Network (Koller and Friedman 2009) structure extraction based approach to extract insights from a combination of traffic sensor data and incident reports. They derive insights that are not obvious to city authorities and present a traffic alert system to deliver these predictions to commuters.

Time Series Based Approaches

If we consider speed and travel time observations as time series observations, our proposition of explaining textual observations using speed and travel time anomalies can be viewed as a time series annotation task. In a related work (Fanaee-T and Gama 2014), variations in the number of bicycles hired at various locations in a city, modeled as time series data, is explained using events in the city such as temporal landmarks, concerts, sports matches, parades, bad weather, and public holidays. Events are detected using location and date specific search queries to a search engine. An ensemble approach is used to build a model that connects major events to the number of bikes being hired. Annotating physiological dynamics of premature babies for risk assessment has been carried out by (Quinn, Williams, and McIntosh 2009), where, variations in physiological observations such as heart rate, blood pressure, and body temperature are modeled using a Switching Linear Dynamical System (SLDS).

Understanding city events using a holistic approach of analyzing both sensor and textual data however has received limited attention. We aim to fill this void by proposing algorithms to relate city events from textual data to anomalies in sensor data. Our approach uses domain knowledge to build and apply multiple linear models to learn normal speed and travel time dynamics.We seek explanations in terms of city events, from both formal (e.g., incident reports) and informal (e.g., twitter) sources, for deviations in traffic dynamics.

Preliminaries

We define representation of city traffic related events and the road network. We also provide a brief overview of Linear Dynamical System (LDS) and propose Restricted Switching Linear Dynamical System (RSLDS) to characterize traffic dynamics.

Traffic Event Representation

We use a quintuple $\langle \hat{e}_t, \hat{e}_l, \hat{e}_{st}, \hat{e}_{et}, \hat{e}_i \rangle$ to represent an event where, $\hat{e}_t$ represents the event type, $\hat{e}_l$ is the location of the event, $\hat{e}_{st}$ is the start time of the event, $\hat{e}_{et}$ is the end time of the event, and $\hat{e}_i$ represents the estimated impact of the event. We use a unified representation of events for both 511.org reported events and traffic events extracted from twitter.

Road Network

The fundamental building block of a road network is called a *link*, represented by l. 511.org provides location information for all the links in San Francisco Bay Area. A road r, is an ordered sequence of links, i.e., $r = [l_1, l_2, ..., l_n]$, where, n is the number of links in road r. The location of a link is specified by start and end lat-long which can be used to reconstruct the road. We collect speed and travel time observations from 511.org for 3,622 links. The whole road network N is a set of roads, $N = \{r_1, r_2, ..., r_m\}$, where m is the total number of roads in the road network.

Linear Dynamical Systems

Time series data with hidden and observed variables naturally occur in the domains such as traffic, healthcare, finance, and system health monitoring. A LDS model (Barber 2012) incorporates both hidden and observed variables as shown in Figure 1(a) with T hidden nodes $\mathrm{h}_{1:T}$ and T observed nodes $\mathrm{s}_{1:T}$ respectively for modeling observations at T time points. A hidden variable captures the state of a system that is not directly observable, e.g., in the context of diseases and symptoms, a disease is hidden while a symptom is observable. In the domain of traffic, the volume of vehicles passing through a link may not be available (511.org does not provide volume data). Further, there may be many other unobserved factors influencing traffic dynamics such as road conditions, visibility, and random effects. These unobserved

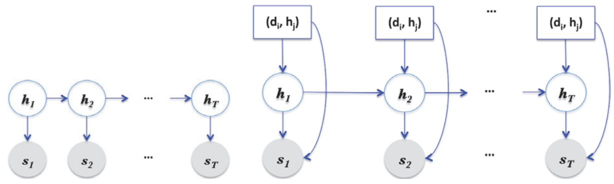

Figure 1: (a) A Linear Dynamical System for T time points with hidden nodes $h_{1:T}$ and observed nodes $s_{1:T}$. (b) A Restricted Switching Linear Dynamical System (RSLDS) with each switch variable indexed by day of week and hour of day (d_i, h_j).

variables at time t may be represented using a hidden node h_t in the LDS model. The average speed of vehicles and average travel time through a link are the observed variables represented using s_t in the LDS model. LDS is formally defined using Equations(1a) and (1b) where $\mathbf{A_t}$ is called the *transition matrix* and $\mathbf{B_t}$ is called the *emission matrix*. $\eta_\mathbf{t}^\mathbf{h}$ and $\eta_\mathbf{t}^\mathbf{s}$ represent the transition and emission noise, respectively.

$$\mathbf{h_t} = \mathbf{A_t h_{t-1}} + \eta_\mathbf{t}^\mathbf{h}, \quad \eta_\mathbf{t}^\mathbf{h} \sim \mathcal{N}(\eta_\mathbf{t}^\mathbf{h}|\bar{\mathbf{h}}_\mathbf{t}, \mathbf{\Sigma_t^h}) \tag{1a}$$

$$\mathbf{s_t} = \mathbf{B_t h_t} + \eta_\mathbf{t}^\mathbf{s}, \quad \eta_\mathbf{t}^\mathbf{s} \sim \mathcal{N}(\eta_\mathbf{t}^\mathbf{s}|\bar{\mathbf{s}}_\mathbf{t}, \mathbf{\Sigma_t^s}) \tag{1b}$$

The hidden state at any time $\mathbf{h_t}$ depends only on the previous hidden state $\mathbf{h_{t-1}}$ (Markovian assumption) and the transition from $\mathbf{h_{t-1}}$ to $\mathbf{h_t}$ is governed by the *transition matrix*. The observation at any time, $\mathbf{s_t}$, depends only on the current hidden state $\mathbf{h_t}$ and is governed by the *emission matrix*. Their joint probability distribution over all the hidden states and observations is given by

$$p(h_{1:T}, s_{1:T}) = p(h_1)p(s_1|h_1) \prod_{t=2}^{T} p(h_t|h_{t-1})p(s_t|h_t) \tag{2}$$

where, the terms $p(h_t|h_{t-1})$ and $p(s_t|h_t)$ are given by

$$p(h_t|h_{t-1}) = \mathcal{N}(h_t|A_t h_{t-1} + \bar{h}_t, \Sigma_t^h) \tag{3a}$$

$$p(s_t|h_t) = \mathcal{N}(s_t|B_t h_t + \bar{s}_t, \Sigma_t^s) \tag{3b}$$

This model offers to capture variation in the transition and emission matrices along with the Gaussian noise. For the domain of traffic, we assume that the transition and emission matrices do not vary over time. Such a model is called a stationary model. Thus, $\mathbf{A_t} \equiv \mathbf{A}$, $\mathbf{B_t} \equiv \mathbf{B}$, $\mathbf{\Sigma_t^h} \equiv \mathbf{\Sigma_h}$, $\mathbf{\Sigma_t^s} \equiv \mathbf{\Sigma_s}$, $\bar{\mathbf{h}}_\mathbf{t} = 0$, and $\bar{\mathbf{s}}_\mathbf{t} = 0$.

The hidden state $\mathbf{h_t}$ is normally distributed with mean $\mathbf{A_t h_{t-1}}$ and covariance Σ_h. The observation $\mathbf{s_t}$ is normally distributed and has mean $\mathbf{B_t h_t}$ and covariance Σ_s.

Problem Formulation

Speed and travel time dynamics in the domain of traffic follows a more or less recurring pattern based on the hour of the day and the day of the week. Traffic dynamics may vary abnormally due to various city traffic related events, varying road conditions, and random effects. We are not guaranteed to have access to all the active city traffic related events and their interactions. A Gaussian Mixture Model (GMM) approach to model speed and travel time variations (Sun, Zhang, and Yu 2006) do not capture the temporal dependencies fundamental to traffic dynamics. Time series techniques such as autoregressive (AR) and autoregressive-integrated-moving-average (ARIMA) models (Lee and Fambro 1999; Moorthy and Ratcliffe 1988) capture temporal dependencies. However, relating later values of speed and travel time with corresponding earlier values alone is not adequate as they cannot capture the latent variables crucial in modeling traffic dynamics. An LDS model offers a better foundation for representing additional factors that are difficult to capture separately in modeling traffic dynamics. The volume of vehicles through a link, associated interactions, and random effects (noise) on traffic dynamics is being approximated by the hidden states ($h_{1:T}$) of the LDS model and the noise terms $\eta_\mathbf{t}^\mathbf{h}$ and $\eta_\mathbf{t}^\mathbf{s}$ as shown in Figure 1(a). The average speed of vehicles passing through a link and average travel time for a link, obtained from sensor data, represents the observed node ($s_{1:T}$) in Figure 1(a).

We broadly categorize various factors that influence traffic into *internal* and *external* factors. *Internal* factors include day of the week, time of the day, and location. *External* factors include city traffic related events such as accidents, breakdowns, music and sporting events. We propose a *Restricted Switching Linear Dynamical System* (RSLDS) as shown in Figure 1(b). We learn one LDS model for each hour of the day and for each day of the week, giving us 24 $\times$ 7 (168) LDS models for each *link*. A switch variable in RSLDS is used to index and select an LDS model on (d_i, h_j), where, d_i is day of week (ranging over 7 days) and h_j is hour of day (ranging over 24 hours). Our approach is similar to Switching Linear Dynamics System (SLDS) (Quinn, Williams, and McIntosh 2009) that allows discrete switches to select an appropriate LDS model. However, SLDS model assumes a Markovian transition between switch configurations, which is violated in the domain of traffic. For example, the *external* factors such as accidents and breakdowns may occur randomly and independently.

Approach

We present our approach to learn models for normal traffic dynamics, tagging anomalies, and utilizing events from textual stream to explain the anomalies below.

Learning Normalcy in Traffic Dynamics

Figure 2 outlines the process of learning normal traffic dynamics. LDS is a linear model and cannot faithfully capture the non-linearity in speed dynamics over time. RSLDS deals with non-linearity by piecewise linear approximation using a collection of LDS models by selecting appropriate linear regime from the collection based on the switch state. Hour of the day has a major influence on traffic dynamics, e.g., morning peak hours (7 am to 9 am) and evening peak hours (5 pm to 7 pm) on a work day typically has slow moving traffic. Day of the week is another important influencer of traffic dynamics, e.g., weekend pattern is different as offices are closed but social, music, or sporting events may occur at certain locations and time durations. We use both the day of the week and the hour of the day to index traffic dynamics and learning normalcy model.

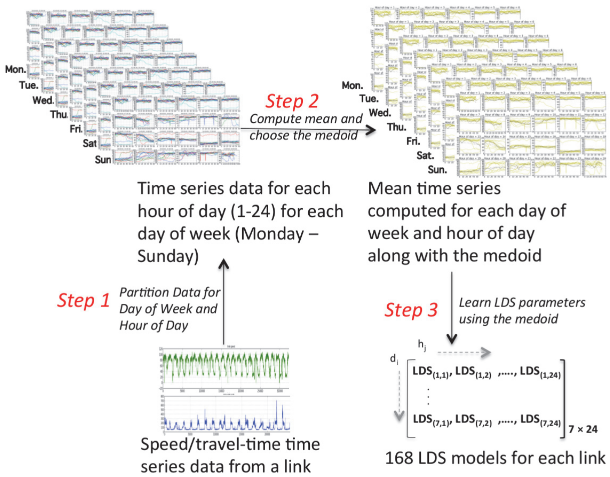

Figure 2: Learning normal traffic dynamics from speed and travel time observations resulting in 168 LDS models.

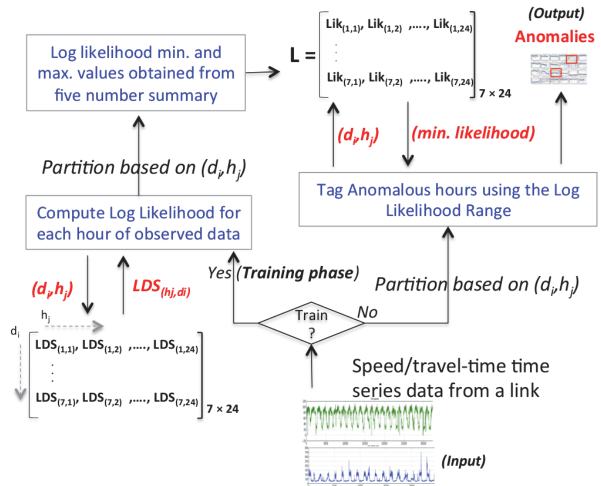

Figure 3: Utilizing 168 LDS models to tag anomalies, which can be tied to a city event reported on textual stream.

Indexing Traffic Dynamics Data In Step 1 of Figure 2, we partition data from a link based on the day of week (Mon-Sun) and further, on the hour of the day (1-24). Hourly speed dynamics for each Monday between May 2014 and Jan 2015, and each hour is shown in Figure 4. Each of the 24 subplots corresponds to the time series of speed variation over each hour of the day. We observe approximate clustering of speed dynamics (light colored lines) in most of the plots, indicating a general hourly trend in speed dynamics. In the first seven hours of the day, starting 12 AM to 7 AM, the average speed of the vehicles remain high and stable, around 80 to 100 km/h. After 7 AM, we observe a decreasing trend in speed until 9 AM, which may be due to morning commute. After an increasing trend in speed around 10 AM, possibly due to subsiding commuter traffic, the speed of vehicles is observed to be stable from 11 AM to 1 PM. A decreasing trend is observed between 1 PM and 2 PM with speeds plummeting to 20 to 30 km/h after 2 PM until 6 PM. This can be due to lunch time and evening rush hour traffic respectively. Closer to 7 PM, peak hour rush subsides resulting in increasing speed trend till 8 PM. After 9 PM, the speed resumes and stabilizes between 80 to 100 km/h.

Selecting Typical Traffic Dynamics In Step 2 of Figure 2, we select typical traffic dynamics by iterating through the index of Step 1. Algorithm 1 describes the selection of a typical traffic dynamics for each hour. The input to Algorithm 1 is the speed observations indexed over *internal factors*. Each hour contains multiple speed sequences $[s_{(m,1)}, \ldots, s_{(m,n)}]$ (if there are five Mondays with 60 observations for an hour, then, m = 1 to 5 and n = 60; speed observations are sampled n times an hour to create each of the m sequences). For computing the average speed at each of the n sampling point, we sum up all the speed values at each sampling index (1 to n) over all the m sequences and divide it by m. Average speed sequence serves as the centroid of all the speed sequences. To select a speed sequence that exists in the real-world (that is, it is *realizable*), we choose the speed sequence that is closest to the centroid using a point-wise Euclidean distance metric, obtaining the medoid.

Algorithm 1 Select medoid for hourly speed plots

Require: Multiple speed observation sequences collected for each (d_i, h_j) where d_i = Monday to Sunday and h_j = 1 to 24, each set containing n speed observations, $[s_{m,1}, \ldots, s_{m,n}]$ where, m indexes over number of speed sequences collected for (d_i, h_j)
Ensure: $[s_1 \ldots, s_n]$ representing the medoid for each $M(d_i, h_j)$
  for each day d from Monday to Sunday **do**
    for each hour h of the day ranging from 1 to 24 **do**
      Select speed values $[s_{(m,1)}, \ldots, s_{(m,n)}]$ from (d_i, h_j)
      Find the average speed $[a_1 \ldots, a_n]$ from m samples
      Select speed sequence $[s_1 \ldots, s_n]$ closest to average
      Set $M(d_i, h_j) = [s_1 \ldots, s_n]$
    end for
  end for

The result of running Algorithm 1 is shown in Figure 4 with mean speed plot (dashed line) and medoid (solid line).

Learning LDS parameters In Step 3 of Figure 2, we learn the parameters of the LDS utilizing the representative traffic dynamics chosen based on Algorithm 1. The LDS parameters are learned for every day of week and every hour of day. LDS is parameterized by $\theta = \{\mathbf{A}, \mathbf{\Sigma_h}, \mathbf{B}, \mathbf{\Sigma_s}, \mu_\pi, \mathbf{\Sigma}_\pi\}$ where $\mathbf{A}$ is the transition matrix, $\mathbf{\Sigma_h}$ is the transition covariance, $\mathbf{B}$ is the emission matrix, $\mathbf{\Sigma_s}$ is the emission covariance, μ_π and $\mathbf{\Sigma}_\pi$ are the mean and covariance of the initial state density $p(h_1)$ in Equation 2. Since the joint distribution of LDS contains hidden variables, Expectation Maximization (EM) algorithm is used for learning the LDS parameters (Ghahramani and Hinton 1996; Barber 2012). From Equation 2, the joint distribution can be rewritten as

$$\ln p(h_{1:T}, s_{1:T}|\theta) = \ln p(h_1|\mu_\pi, \mathbf{\Sigma}_\pi) \; + \sum_{t=2}^{T} \ln p(h_t|h_{t-1}, \mathbf{A}, \mathbf{\Sigma_h}) + \sum_{t=1}^{T} \ln p(s_t|h_t, \mathbf{B}, \mathbf{\Sigma_s}) \tag{4}$$

Equation 4 has explicit parameterization and represents the log likelihood of data given the parameters. EM algorithm chooses the initial parameters θ^{old} and evaluates $p(h_{1:T}|s_{1:T}, \theta^{old})$ in the expectation step. In the maximization step, the expectation of the log likelihood function represented by $\mathbb{E}_{h_{1:T}|\theta^{old}}[\ln p(h_{1:T}, s_{1:T}|\theta)]$ is maxi-

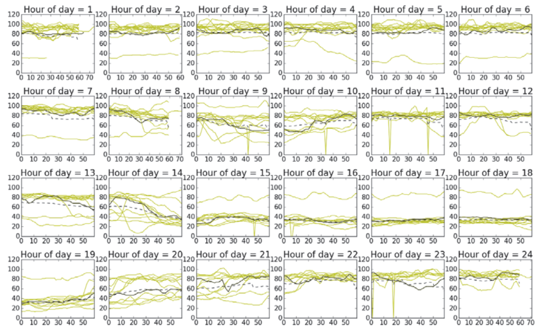

Figure 4: Hourly plot of speed variations over time for all the Mondays from May 2015-June 2015 for a link.

mized with respect to θ. The parameters are updated to $\mathbf{A}^{\mathbf{new}}, \mathbf{\Sigma}_{\mathbf{h}}^{\mathbf{new}}, \mathbf{B}^{\mathbf{new}}, \mathbf{\Sigma}_{\mathbf{s}}^{\mathbf{new}}, \mu_{\pi}^{\mathbf{new}}$, and $\mathbf{\Sigma}_{\pi}^{\mathbf{new}}$. After maximization step, if the convergence criteria is not satisfied, the new parameter setting for LDS, θ^{new} replaces θ^{old} and the algorithm returns to the expectation step.

Detecting Anomalous Traffic Dynamics

Figure 3 outlines the process of learning normalcy in terms of log likelihood score and utilizing it to tag anomalies. In the training phase, we utilize the 168 LDS models learned for each link indexed by *internal factors*, $\theta(d_i, h_j)$, for estimating the log likelihoods, $\ln p(h_{1:T}, s_{1:T}|\theta(d_i, h_j))$ as shown in Equation (4). In the training phase, we learn the typical likelihood values after aggregating the log likelihood scores for the entire dataset partitioned by (d_i, h_j), thereby capturing normalcy. We utilize a non-parametric approach of five number summary (*minimum*, first quartile, median, third quartile, and *maximum*) over the log likelihood scores for each partition indexed by (d_i, h_j). The log likelihood range (*minimum* and *maximum*) exists for each day of the week and the hour of the day, (d_i, h_j). In the testing phase, we compute the log likelihood score for the observed data using the appropriate LDS model $\theta(d_i, h_j)$. If the log likelihood value for a particular day of week and the hour of the day is less than the *minimum* log likelihood value (retrieved from matrix $\mathbf{L}$ in Figure 3), we tag the traffic dynamics as anomalous.

Traffic Events for Explaining Anomalies

For every city traffic event collected from textual data, we detect anomalies in traffic pattern as outlined in Figure 5. We examine the city traffic events using their location, start time, and end time. Based on the location of the event, we select links within a radius r km from the event. We run the anomaly detection step on the temporally selected data points from these links. If an event from the textual stream has a corresponding anomaly in the link data, we hypothesize that the event *explains* the anomaly and the anomaly is *explained by* the event. Algorithm 2 determines textual events *E*, that explains anomalies in sensor data. The radius r is an input parameter which can be changed but it is set to 0.5 km in our experiments. The adjusted duration of an event, $\Delta t_e = (\hat{e}_{st} - h, \hat{e}_{et} + h)$, where, h is set to 1 hour

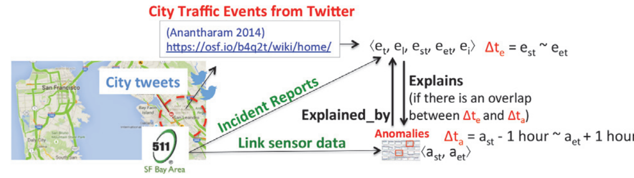

Figure 5: City traffic events (textual data) used to explain anomalies in traffic dynamics (sensor data).

(lowest granularity of our analysis), is used to select sensor data from all the links within the radius of r km from the event location, $\hat{e}_l$. If the selected link data has anomalies, possibly explained by the textual event, then the event is accumulated in $E_{explained}$ as shown in Algorithm 2.

Algorithm 2 Explaining Traffic Events by Anomalies

Require: Set of city traffic events E containing event tuples $\langle \hat{e}_t, \hat{e}_l, \hat{e}_{st}, \hat{e}_{et}, \hat{e}_i \rangle$, latitude and longitude of all the 3,622 links in the road network, log likelihood range matrix $\mathbf{L}$ indexed by (d_i, h_j), radius parameter r km to select the links, and time parameter h for adjusting event duration

Ensure: $E_{explained}$ containing all events with corresponding anomalies in sensor data

for each event quintuple $\langle \hat{e}_t, \hat{e}_l, \hat{e}_{st}, \hat{e}_{et}, \hat{e}_i \rangle$ in **E do**
  Find hourly time range $\Delta t_e = (\hat{e}_{st} - h, \hat{e}_{et} + h)$
  Let M be all the links within the radius of r km from the event location $\hat{e}_l$
  for each link $l \in M$ **do**
    Select data for link l filtered by duration Δt_e
    $\mathrm{H}(d_i, h_j) \leftarrow$ Compute the hourly log likelihood on selected data using Equation (4)
    if hourly log likelihood $\mathrm{H}(d_i, h_j) <$ minimum log likelihood from $\mathbf{L}(d_i, h_j)$ **then**
      $E_{explained} \leftarrow \langle \hat{e}_t, \hat{e}_l, \hat{e}_{st}, \hat{e}_{et}, \hat{e}_i \rangle$
    end if
  end for
end for

Evaluation

We conducted a large scale evaluation of our approach on real-world traffic sensor and twitter data collected for a year.

Traffic Dataset from 511.org and Tweets: We collected 1,638 city traffic related events from 511.org and we extracted 39,208 city traffic events from over 20 million tweets collected from May 2014 to May 2015 for San Francisco Bay Area, utilizing an openly available city traffic event extraction tool (Anantharam 2014), resulting in a total of 40,846 city traffic events. 511.org also provides minute by minute speed and link travel time data for 3,622 links resulting in over 1.4 billion time series data points. Out of 3,622 links, 1,088 links do not have any data points for the entire year. Further, there are partially missing data points in the time series for the remaining 2,534 links.

Evaluation Strategy: Traffic events from 511.org are reliable since it is reported by city authorities. We use these events as a reference in our evaluation. Algorithm 2 iterates over 1,638 events from 511.org to explain anomalies in traffic. We evaluate Algorithm 2 for finding 511.org event manifestations in sensor data. Further, we extend the evaluation to 39,208 events extracted from twitter and report our results.

Evaluation over 511.org Traffic events: We evaluate our approach by analyzing the co-occurrence of the event in textual data with the anomaly detected in the sensor data. Table 1 presents the evaluation summary for all the 1,638 511.org events and 39,208 twitter traffic events. Events with no links near them (for r = 0.5 km) are placed under *No*

Source	Total Events	No Links	Missing Data	No Anomalies	Anomalies
511.org	1,638	201	901	145	391
Twitter	39,208	36,436	1942	18	812

Table 1: Evaluation results for all the events using Algorithm 2 with parameter setting: h = 1 hour and r = 0.5 km.

Links. If there are links near an event, with data that may be missing (for the event duration), then they are characterized as *Missing Data*. Links near an event with data are used to tag anomalies and the result is placed under *No Anomalies* and *Anomalies*. We call the events corroborated with anomalies in any of the link sensor data near the event as being *explained*. Events without accompanying anomalies in any of the link sensor data are called *un-explained*. For a palatable comparison, we present percentages of events from 511.org and twitter that explain anomalies in Figure 6. We observe a larger set of links near events from 511.org relative to twitter events as shown in Figure 6 (bottom). Out of 33% 511.org events with complete sensor data, we could explain 72% of them. Figure 6 (top) presents a sample output of Algorithm 2 after processing 10 events. For events marked in bold, we found anomalies in traffic dynamics possibly explained by the event with the following insights: (a) Long-term events may not manifest as anomalies in sensor data. RSLDS normalcy model is trained over the entire year of average speed and travel time observations. Long-term events such as construction activities may span several months. Since the data we have is for a year, such long-term events are part of the normalcy model and may not be tagged anomalous as observed in Figure 6 (top). Events, such as accidents and disabled vehicles, are short lived events that may manifest as anomalous traffic. (b) Location and start time of the event may impact its manifestation in sensor data. Events near crowded places would most likely manifest as anomalies. Events occurring during off-peak hours are less likely to manifest in sensor data compared to the events occurring during peak-hours. (c) Missing data creates challenges for associating anomalies with events. Among the 2,534 links with data, there is missing data for many days in a year due to maintenance and sensor failures resulting in decreased coverage.

Evaluation over Twitter Traffic Events: Twitter traffic events are dispersed widely across the city resulting in reduced or missing links near many events. There are 36,436 twitter traffic events with no links near them as shown in Table 1 due to significantly lower sensor data coverage. Consequently, we observe that the coverage can be significantly improved by augmenting information from sensor data with that from twitter events as shown in Figure 6 (bottom). We expected a higher twitter traffic event manifestation in sensor data since people will most likely report events of significant impact while 511.org reports all possible traffic related events that may have varying impact. Out of 2% twitter traffic events with complete sensor data, we could corroborate 97% of it with anomalies.

Scalability Challenges: There are 2,534 links with data. For each link, we learn 168 LDS models by analyzing over 1.4 billion data points resulting in a total of 425,712 (= 2,534 $\times$ 168) LDS models. The size of the traffic dataset is around 30 GB. Learning LDS parameters and the criteria for anomaly is computationally expensive. For each link with one year of data, we estimated 25 minutes for learning LDS models and 15 minutes for computing the criteria for anomaly, resulting in a total processing time of 40 minutes per link. Extrapolating processing time for all the links, we get 1,689 hours ($\frac{40minutes \times 2{,}534}{60minutes}$) ($\approx$2 months). Initial processing was done with 2.66 GHz, Intel Core 2 Duo processor with 8 GB main memory. We then exploited inherent "embarrassing parallelism" to devise a scalable implementation of our approach on Apache Spark (Zaharia et al. 2010) that takes less than a day. The Apache Spark cluster used in our evaluation has 864 cores and 17TB main memory.

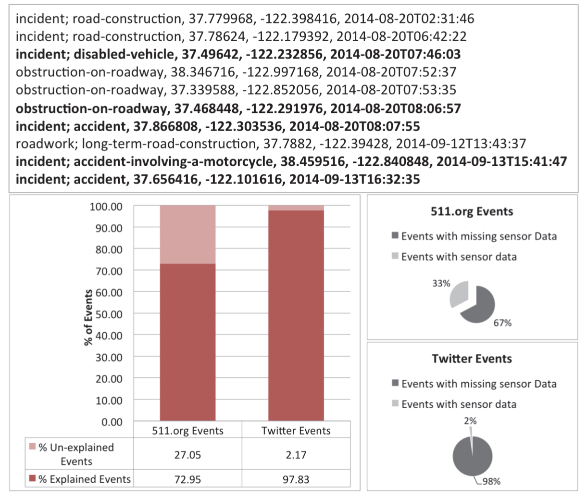

Figure 6: City traffic events with available and missing link data and percentage of events explained by Algorithm 2

Conclusion and Future Work

Normal traffic dynamics can be captured using RSLDS, a variant of LDS model, that utilizes domain knowledge to segment nonlinear traffic dynamics into linear components. Utilizing the normalcy model, we could explain anomalies in traffic sensor data using traffic events from textual data. We could also associate real-world events that impact traffic by determining anomalies in traffic pattern. Further, a large scale evaluation of our approach on a real-world dataset collected for a year corroborated 72% of 511.org events and 97% of twitter traffic events in terms of anomalous traffic dynamics. As a future work, RSLDS model capturing temporal dynamics can be utilized to study various traffic event types and associated speed and travel time dynamics and predict traffic dynamics based on the traffic events from textual streams.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. EAR 1520870 titled "Hazards SEES: Social and Physical Sensing Enabled Decision Support for Disaster Management and Response". Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We would also like to acknowledge the EU FP7 Citypulse project contract number: 609035 for providing a broader motivation for this work.

References

Anantharam, P. 2014. Extracting city traffic events from social streams. https://osf.io/b4q2t/wiki/home/.

Anderson, J., and Bell, M. 1997. Travel time estimation in urban road networks. In *Intelligent Transportation System, 1997. ITSC'97., IEEE Conference on*, 924–929. IEEE.

Barber, D. 2012. *Bayesian reasoning and machine learning*. Cambridge University Press.

Bishop, C. M. 2006. *Pattern recognition and machine learning*. springer.

Boulos, M. N. K.; Resch, B.; Crowley, D. N.; Breslin, J. G.; Sohn, G.; Burtner, R.; Pike, W. A.; Jezierski, E.; and Chuang, K.-Y. S. 2011. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, ogc standards and application examples. *International journal of health geographics* 10(1):67.

Chawla, S.; Zheng, Y.; and Hu, J. 2012. Inferring the root cause in road traffic anomalies. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, 141–150. IEEE.

Crooks, A.; Croitoru, A.; Stefanidis, A.; and Radzikowski, J. 2013. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS* 17(1):124–147.

Daly, E. M.; Lecue, F.; and Bicer, V. 2013. Westland row why so slow?: fusing social media and linked data sources for understanding real-time traffic conditions. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, 203–212. ACM.

De Fabritiis, C.; Ragona, R.; and Valenti, G. 2008. Traffic estimation and prediction based on real time floating car data. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, 197–203. IEEE.

Fanaee-T, H., and Gama, J. 2014. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence* 2(2-3):113–127.

Ghahramani, Z., and Hinton, G. E. 1996. Parameter estimation for linear dynamical systems. Technical report, Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Goodchild, M. F. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69(4):211–221.

Horvitz, E.; Apacible, J.; Sarin, R.; and Liao, L. 2012. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. *arXiv preprint arXiv:1207.1352*.

IBM. 2014. Smarter traffic. http://www.ibm.com/smarterplanet/us/en/traffic_congestion/ideas/.

Ko, J., and Guensler, R. L. 2005. Characterization of congestion based on speed distribution: a statistical approach using gaussian mixture model. In *Transportation Research Board Annual Meeting*. Citeseer.

Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Lécué, F.; Schumann, A.; and Sbodio, M. L. 2012. Applying semantic web technologies for diagnosing road traffic congestions. In *The Semantic Web–ISWC 2012*. Springer. 114–130.

Lee, S., and Fambro, D. 1999. Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting. *Transportation Research Record: Journal of the Transportation Research Board* (1678):179–188.

Lee, W.-H.; Tseng, S.-S.; and Tsai, S.-H. 2009. A knowledge based real-time travel time prediction system for urban network. *Expert Systems with Applications* 36(3):4239–4247.

Moorthy, C., and Ratcliffe, B. 1988. Short term traffic forecasting using time series methods. *Transportation planning and technology* 12(1):45–56.

Nagarajan, M.; Sheth, A.; and Velmurugan, S. 2011. Citizen sensor data mining, social media analytics and development centric web applications. In *Proceedings of the 20th international conference companion on World wide web*, 289–290. ACM.

Pattara-Atikom, W.; Pongpaibool, P.; and Thajchayapong, S. 2006. Estimating road traffic congestion using vehicle velocity. In *ITS Telecommunications Proceedings, 2006 6th International Conference on*, 1001–1004. IEEE.

Quinn, J. A.; Williams, C. K.; and McIntosh, N. 2009. Factorial switching linear dynamical systems applied to physiological condition monitoring. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31(9):1537–1551.

Rajkumar, R. R.; Lee, I.; Sha, L.; and Stankovic, J. 2010. Cyber-physical systems: the next computing revolution. In *Proceedings of the 47th Design Automation Conference*, 731–736. ACM.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, 851–860. ACM.

Sheth, A.; Anantharam, P.; and Henson, C. 2013. Physical-cyber-social computing: An early 21st century approach. *Intelligent Systems, IEEE* 28(1):78–82.

Sun, S.; Zhang, C.; and Yu, G. 2006. A bayesian network approach to traffic flow forecasting. *Intelligent Transportation Systems, IEEE Transactions on* 7(1):124–132.

Wanichayapong, N.; Pruthipunyaskul, W.; Pattara-Atikom, W.; and Chaovalit, P. 2011. Social-based traffic information extraction and classification. In *ITS Telecommunications (ITST), 2011 11th International Conference on*, 107–112. IEEE.

Zaharia, M.; Chowdhury, M.; Franklin, M. J.; Shenker, S.; and Stoica, I. 2010. Spark: Cluster computing with working sets. In *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*, HotCloud'10, 10–10. Berkeley, CA, USA: USENIX Association.

Zeng, K.; Liu, W.; Wang, X.; and Chen, S. 2013. Traffic congestion and social media in china. *IEEE Intelligent Systems* 28(1):72–77.