Traffic Updates: Saying a Lot While Revealing a Little

John Krumm and Eric Horvitz
Microsoft Research
Redmond, Washington USA
jckrumm@microsoft.com, horvitz@microsoft.com

Abstract

Taking speed reports from vehicles is a proven, inexpensive way of inferring traffic conditions. However, due to concerns about privacy and bandwidth, not every vehicle occupant may want to transmit data about their location and speed in real time. We show how to drastically reduce the number of transmissions in two ways, both based on a Markov random field for modeling traffic speed and flow. First, we show that a only a small number of vehicles need to report from each location. We give a simple, probabilistic method that lets a group of vehicles decide on which subset will transmit a report, preserving privacy by coordinating without any communication. The second approach computes the potential value of any location's speed report, emphasizing those reports that will most affect the overall speed inferences, and omitting those that contribute little value. Both methods significantly reduce the amount of communication necessary for accurate speed inferences on a road network.

1 Introduction

Accurate traffic speeds are vital for computing efficient routes for vehicles on roads. Dedicated road sensors, such as embedded induction loops, are expensive to install and maintain. An obvious alternative is to use the mobile phones of vehicle occupants to sense and report speeds to a central server. As of early 2018, 95% of Americans owned a cellphone of some kind, and 77% owned a smartphone (Center 2018). Both fractions are rising. As an example, Waze gathers GPS data from its users at a rate of one reading per second to compute road speeds to use for routing (Parmy Olson 2014). These GPS data not only uses bandwidth, but it is vulnerable to privacy attacks. In fact, researchers were able to create fake Waze accounts that allowed them to track Waze users (**?**), a vulnerability that was patched by Waze (Waze 2016). Even coordinating among vehicles with vehicle-to-vehicle communication has negative privacy implications (Williams 2017).

Problems with bandwidth and privacy can be fixed with advanced compression (Lelewer and Hirschberg 1987) and location privacy techniques (Krumm 2009), respectively. However, another solution is to transmit less data. This paper introduces two techniques to reduce the number of revealing transmissions required to maintain accurate inferences about traffic speeds. Both approaches are based on a Markov random field (MRF) that models varying quantities of vehicle speed report data as well as the observed correlations between traffic conditions on different parts of the road network. Using this model, we show that a small number of speed reports can be combined to infer traffic conditions on all the roads in the network.

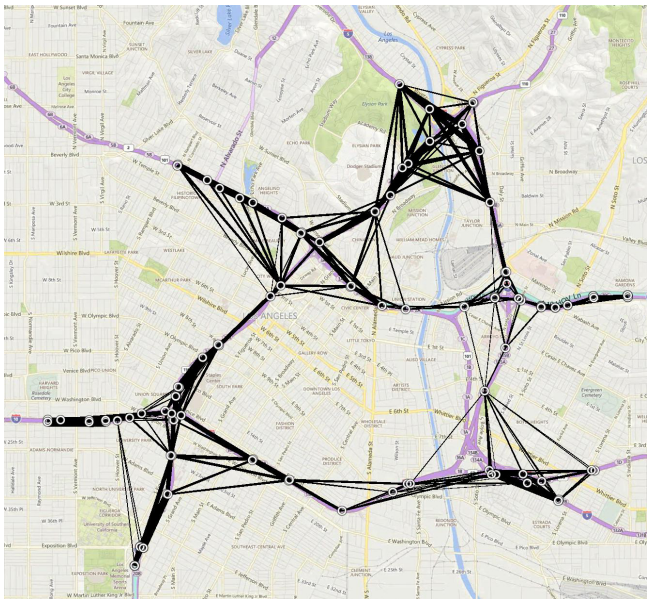

Figure 1: The white circles show 109 traffic measurement stations around central Los Angeles, California, USA. The black lines show which stations were connected in the Markov random field.

Our first approach to reduce privacy-compromising transmissions from vehicles is to show that only a relatively small number of vehicles needs to report from any road segment, and that we can choose the subset of vehicles to report without any communication. Normally this coordination might be accomplished with each vehicle communicating to a central server or with with vehicle-to-vehicle communication. Instead, we develop a probabilistic technique where each vehicle makes an independent decision to transmit, while still

maintaining approximately the desired number of vehicle reports.

Our second technique is sensitive to the fact that not all road segments need a speed measurement. In fact, if traffic is moving normally, no reports are necessary. We formalize this with a value-of-information analysis. Each vehicle runs a local version of the MRF to assess approximately how much its own speed report would affect traffic inferences on the whole road network. Reports that would have a larger effect on the network are prioritized over those that would have little or no effect.

Next we describe the traffic data and MRF model that we used for our experiments.

2 Traffic Data

Our experimental data consists of freeway speeds and flows from the state of California in the USA. Traffic on roads is often characterized by its speed and flow, where flow indicates the number of vehicles passing a certain point in a given amount of time. Such data is available free via the California Performance Measurement System (PeMS), which provides a wide variety of real-time and historical data for freeways in California (of Transportation (Caltrans) 2018). PeMS includes traffic data collected from over 35,000 traffic detectors that report every 30 seconds. The main type of detector is inductive loops, but there are also side-fire radars and magnetometers. PeMS aggregates this data into reports from discretely located measurement stations, each of which covers all freeway lanes in the same direction. For instance, one station may pertain to all northbound lanes at a particular location on Interstate 5. PeMS reports the data at five-minute intervals, giving the mean speed and count of vehicles per five minutes (flow) at each station. We maintain this five-minute time discretization throughout our analysis.

For our experiments, we chose data from 109 different freeway measurement stations inside in a 3.1 mile (5 kilometer) radius around the center of Los Angeles, California, USA, shown in Figure 1. Figure 2 shows the speed and flow averaged over all 109 stations for one day. In the subsequent sections of this paper, we refer to these measurement stations as simply "stations". We used data from the first six months of 2017 for training and the last three months for testing. We used months 7-9 for value-of-information training described in Section 7.

Next we describe a method to infer speeds at all the stations based on noisy speed reports from only a subset of the stations.

3 Markov Random Field for Traffic

We develop a Markov random field (MRF) for statistical modeling of macroscopic traffic speeds and flows. This model is particularly well-suited to our task of reducing the number of of traffic speed reports, because it propagates measurements from a subset of the stations to make inferences for all the stations

3.1 Previous Work

Previous traffic models have included the vector autoregressive (VAR) model of Liu et al., which models speeds on a particular road segment as a weighted, linear sum of speeds on other road segments (Liu et al. 2016). Hongzi Zhu et al. use a multi-channel singular spectrum analysis (MSSA) model to infer traffic speeds despite noise and missing values (Zhu et al. 2009). In the work by Yanmin Zhu et al., the authors find correlations between traffic on different roads through principal component analysis (PCA) (Zhu et al. 2013). JamBayes uses a Bayesian network to infer and predict traffic based on several features like current traffic, road incidents, weather, holidays, and planned events (Horvitz et al. 2012). Zhang et al. demonstrate a deep, residual network to predict crowd flows, including inputs such as weather (Zhang, Zheng, and Qi 2017). There are also microscopic traffic flow models, often based on physics (e.g. (Nagatani 2002)), but our interest here is in a more data-driven model that can be used to understand and exploit the value of small amounts of traffic data for making network-wide inferences.

An MRF model has been applied to traffic before, e.g. for modeling images of vehicles at intersections (Kamijo et al. 2000) and for using Twitter to sense traffic (Chen, Chen, and Qian 2014). The most relevant previous MRF work is that of Kataoka et al. (Kataoka et al. 2014). Their MRF helps fill in missing traffic data based on sensed data at other locations. The distributions in their MRF are parametric, while we use nonparametric probability distributions learned from historical data. Our nonparametric formulation has the advantage of representing arbitrary relationships between traffic conditions on different roads. This more flexible representation allows us to model both the speed and flow, rather than just density. Furthermore, we use multiple MRFs targeted at different days of the week and times of day, while Kataoka et al's approach uses a single MRF for all time. Hu et al. also use an MRF for traffic estimation, specifically looking for certain "seed roads" which are most indicative of traffic conditions on other roads (Hu et al. 2016). Their model produces a binary indicator that shows if the traffic is moving faster or slower than average.

The main innovation of our MRF model, however, is how we use it. The VAR, MSSA, PCA, and MRF models referenced above are mostly aimed at filling in missing traffic values by exploiting correlations discovered in historic traffic data. In contrast, our work aims at using a small number of traffic reports, including possibly zero, while still maintaining accuracy.

3.2 Markov Random Field

Our MRF model represents each traffic variable (i.e. speed and flow) in a road network as a connected node in a graphical model. For each of the N measurement stations, the scalar variables s_i and f_i, $i = 1 \dots N$, represent the mean speed and flow, respectively. We measure speed in mph and flow as the number of vehicles passing a point in five minutes. We consider these variables as unknown, but they can be inferred from noisy speed measurements made by passing vehicles. We represent such a measurement as $\hat{s}_i$, where

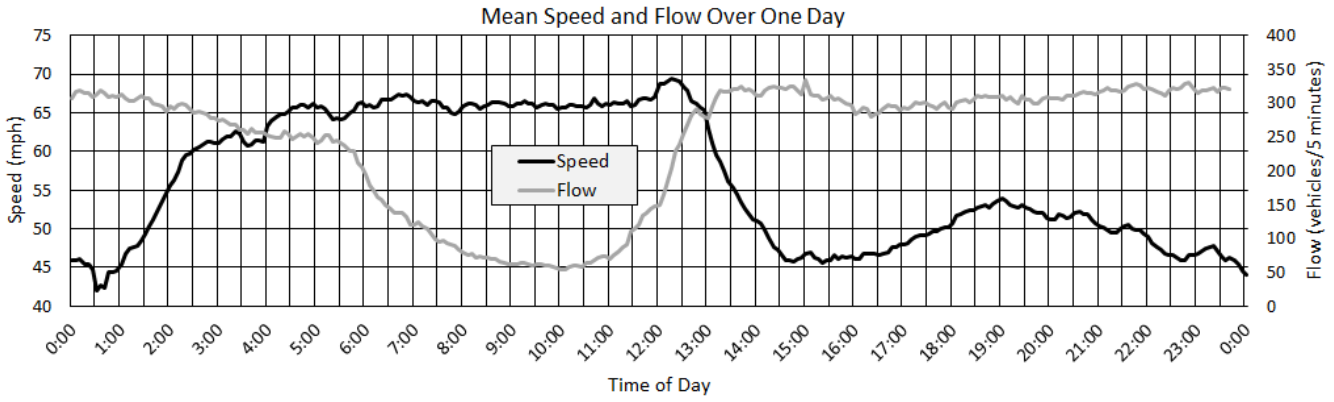

Figure 2: This shows the mean speed and flow over all 109 measurement stations around downtown Los Angeles, California, USA on 19 April 2017.

the hat indicates this is a noisy measurement that likely deviates from the actual value. It would be difficult for a single probe vehicle to measure flow, so we assume a vehicle can only report speed.

The MRF depends on so-called compatibility functions between variables. These are often expressed as pairwise joint probability functions between pairs of variables. We have the following discrete, joint PDFs to represent relationships among the actual speeds, actuals flows, and measured speeds:

- $P_{s_i,s_j}(s_i, s_j)$ is the relationship between speeds at measurement stations i and j. The most useful of these is when $i \neq j$, which represents the joint PDF of speeds at two different stations.

- $P_{s_i,f_i}(s_i, f_i)$ is the relationship between speed and flow at station i. Although there are approximations of this relationship (e.g. (Akcelik 1996)), we learn the relationship from the data.

- $P_{s_i,\hat{s}_i}(s_i, \hat{s}_i)$ is the relationship between the actual and measured speeds at station i. These are different due to measurement noise of the vehicle. Since $\hat{s}_i$ is a given measurement, we abbreviate this PDF with the simpler $P_{s_i}(s_i)$ and call it a measurement distribution.

Graphically, we can think of the MRF as an undirected graph with a node for each variable and an edge for each joint PDF, as in Figure 3. Seeing it this way gives rise to ideas for other topologies, such as introducing joint PDFs between the roads' flow variables or between one road's speed an another road's flow. While the MRF offers this flexibility, we found the given topology to be adequate for our purposes.

Adopting the development in (Yedidia, Freeman, and Weiss 2003), the overall joint probability of the speeds, flows, and measured speeds of the measurement stations is

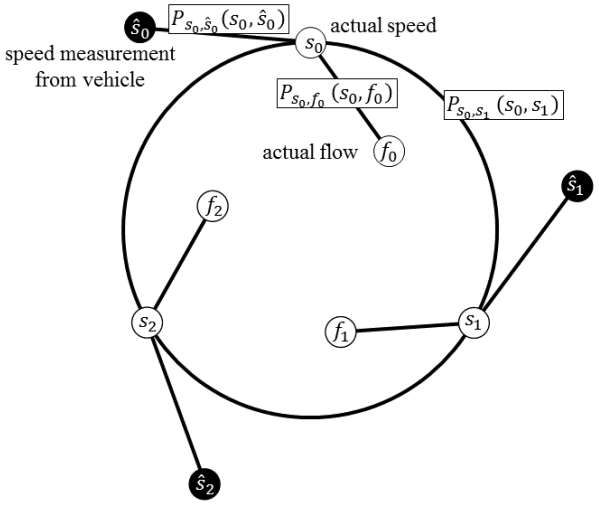

Figure 3: This is a graphical representation of the variables for a road network with three measurement stations. The measured variables are associated with the filled-in circles, and the unknowns are in the white circles.

$$P(\{s\},\{f\},\{\hat{s}\}) = \prod_{(ij)} P_{s_i,s_j}(s_i, s_j) \times \prod_i P_{s_i,f_i}(s_i, f_i) \times \prod_i P_{s_i,\hat{s}_i}(s_i, \hat{s}_i) \quad (1)$$

Here $\{s\}$ is the set of all speeds, and similarly for $\{f\}$. The term $\{\hat{s}\}$ is the set of measured speeds. The notation (ij) indicates all possible distinct, unordered pairs of i and j.

Intuitively, for a given set of measurements $\{\hat{s}\}$, we want to find values of s and f that maximize the joint probability

in Equation (1). This gives the inferred speeds and flows that we want.

3.3 Loopy Belief Propagation

The joint distribution in Equation 1 is difficult to optimize in a straightforward way, given that we have over 100 measurement stations, each with its own speed s_i, flow f_i, and a potential measurement $\hat{s}_i$. To make this easier, we begin with a slight reformulation of Equation 1 which does not include the inviolate speed measurements $\hat{s}_i$, because they do not change. This means the joint PDFs for measurements become simply one-dimensional PDFs over the speed values. That is, $P_{s_i,\hat{s}_i}(s_i, \hat{s}_i)$ becomes simply $P_{s_i}(s_i)$, which gives the distribution of speeds based on a tangible speed measurement at station i. The joint probability with only the unknown variables is

$$P(\{s\}, \{f\}) = \prod_{(ij)} P_{s_i,s_j}(s_i, s_j) \prod_i P_{s_i,f_i}(s_i, f_i) P_{s_i}(s_i) \tag{2}$$

Belief propagation (BP) is an algorithm that can find the *a posteriori* distributions, or beliefs, of all the unknown variables in a joint PDF like Equation 2 (Yedidia, Freeman, and Weiss 2003). That it finds the distributions, rather than simply the maximum *a posteriori*, is advantageous, because the distributions represent the uncertainty of the inferences. BP proceeds by passing messages along the edges of the joint PDF graph. The details of the messages are available in many tutorials, including (Yedidia, Freeman, and Weiss 2003). Each node receives messages from all its connected nodes. A received message is a distribution over the recipient's domain of possible values that gives the sender's belief of what the recipient's state should be. This message is computed from all the sender's connected nodes except for the intended recipient node. The recipient combines the messages to form its *a posteriori* distribution. In a graph without loops, these messages need to be passed only once to provably converge to the exact *a posteriori* distributions.

In our problem, the graph has loops, as shown in the example graph in 3. Fortunately, repeatedly sending updated messages, i.e. loopy BP, often works in these cases, with the *a posteriori* distributions converging after enough rounds of message-passing. We stopped our program's message passing after the mean absolute difference in all messages dropped below 0.1 or if the number of messaging iterations exceeded 100. For our experiments, loopy BP always converged. The result of loopy BP is a PDF of the inferred values of speed and flow at each station. The PDF represents the uncertainty of the estimate. We used the mode of the PDFs to extract inferred scalars to compare to ground truth. In Section 7 we need individual vehicles to run loopy BP. On a conventional desktop PC, our custom BP code converged in an average of 20 milliseconds, so it is feasible to run in a vehicle.

3.4 Joint Probabilities for Road Speed and Flow

The joint probabilities $P_{s_i,s_j}(s_i, s_j)$ and $P_{s_i,f_i}(s_i, f_i)$ describe how speeds and flows on the road network vary with each other, and they come from temporally co-occurring pairs of speed and flow in our six months of freeway training data. We computed these joint probabilities in the normal way by normalizing frequency counts of discretized pairs of measurements. For all our MRF inferences, we discretized speeds into 5 mph bins, and we discretized flows into 25 vehicles/5 minutes bins. For the speed-flow probabilities $P_{s_i,f_i}(s_i, f_i)$, we computed a joint PDF for each measurement station. The speed-speed PDFs, $P_{s_i,s_j}(s_i, s_j)$, relate speeds between pairs of measurement stations. To limit the complexity of our MRF, we only computed a speed-speed PDF between station i and j if station j was one of station i's ten nearest neighbors or vice-versa, using the great circle distance. This is based on an assumption that traffic effects are mostly local, which was justified by the overall accuracy of our model. For pairs of measurement stations not in each other's set of nearest neighbors, there was no edge between them in the MRF. The black lines in Figures 1 and 3 show which pairs of stations were connected by a joint PDF.

To account for possibly different relationships between traffic on different days of the week and at different times of day, we split a canonical week into five-minute intervals and computed a separate set of joint PDFs, and thus a separate MRF, for each five-minute interval. For example, we had one MRF for Mondays from 8:00 a.m. - 8:05 a.m. and a different MRF for Saturdays from 10:45 p.m. - 10:50 p.m., giving a total of 2016 MRFs to cover one representative week.

3.5 Probabilities for Measured Speed

Practically, representing the uncertainty in speed measurements is important in that it allows the MRF to settle on speed inferences that are a compromise between the speed measurements and the joint PDFs. The joint PDFs serve as a sort of prior on speeds, flows, and their relationships. If measured speeds were injected into the MRF with no uncertainty, loopy belief propagation would simply converge to the measured speeds, ignoring the joint PDFs.

One important representational advantage of the MRF becomes apparent with the measured speed distributions. In our algorithm for selecting which vehicles should report their speeds, some measurement stations have no associated speed report. We represent this $P_{s_i}(s_i)$ as simply a uniform speed distribution over the range zero to the maximum speed observed at the station during training. This essentially lets the associated node float based solely on messages passed in from its connected nodes. One extreme we explore in our experiments is to set *all* speed reports to uniform and let the entire MRF float to a set of inferences which are essentially independent of any measurements.

It is also important to fairly represent the uncertainty of speed reports from vehicles. We explore this in the next section.

4 Probabilistic Speed Reports from GPS

In our scenario, a subset of vehicles on the road report their speeds to a central server, which in turn makes network-wide speed and flow estimates using an MRF. This section describes how we model the inherent noise in these speed

reports, due to both the natural variation in speed among a group of vehicles and due to measurement noise.

Our freeway traffic data gives mean speeds every five minutes at each measurement station. However in our testing, we need to model speed reports from individual vehicles. Here we seek to derive a probability distribution that describes these invidual reports. We model the variation from two sources: the natural differences between vehicle speeds traveling on a road and the noise due to GPS measurement error.

The first source of speed variation is the natural differences of speeds among vehicles traversing the same road segment. This variation has been studied for its effects on crashes (Kockelman and Ma 2010) and audible noise (Iannone, Guarnaccia, and Quartieri 2013). While our traffic data gives mean speeds, we want to assess the effect of speed reports from different vehicles passing the same measurement point on the road, whose speeds will inevitably be different. In (Iannone, Guarnaccia, and Quartieri 2013), Iannone et al. review work on speed variation, noting that the dominant model is a simple Gaussian probability distribution, which works best for free-flowing traffic. (Iannone, Guarnaccia, and Quartieri 2013) says that the standard deviation $\sigma_{s_{natural}}$ normally varies between 3.1 and 12.5 mph. Using PeMS data sampled at 30 seconds and aggregated over 5-minute periods, we found a mean $\sigma_{s_{natural}}$ of 5.3 mph. Using regression, we also found that the road's mean speed and flow were poor predictors of $\sigma_{s_{natural}}$, so we chose to use this constant value throughout our analysis instead of trying to infer it based on road parameters or traffic measurements.

The second source of speed variation is measurement noise. We assume that vehicles measure their speed using pairs of location measurements from their onboard GPS sensors [1]. Using the common Gaussian noise assumption for GPS (Diggelen 2007), a location measurement vector is distributed as $\mathbf{x}_i \sim \mathcal{N}([x_i, y_i]^T, \sigma_g^2 I)$. We approximate the standard deviation of GPS as σ_g = 3 meters. Taking the vector difference of two measurements gives the velocity vector:

$$\mathbf{v}_i = \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i-1}}{\Delta t_i}$$

where $\Delta t_i = t_i - t_{i-1}$, $\boldsymbol{\mu}_i = [x_i, y_i]^T$, and $\boldsymbol{\mu}_{i-1} = [x_{i-1}, y_{i-1}]^T$. For our simulation, we take $\Delta t_i = 2$ seconds.

If the two location location measurements are independent, their variances will add, and the two-dimensional distribution of the velocity vector will be:

$$\mathbf{v}_i \sim \mathcal{N}\left(\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i-1}}{\Delta t_i}, 2\left(\frac{\sigma_l}{\Delta t_i}\right)^2 I\right)$$

where I is the 2x2 identity matrix.

We now have a distribution for the velocity vector, but we are ultimately interested in the distribution for scalar speed, which is the magnitude of velocity. For the case of a bivariate normal with a diagonal covariance matrix, the distribution of the magnitude follows a Rician distribution (Rice 1945):

$$||\mathbf{v}_i|| \sim \text{Rice}\left(\frac{||\boldsymbol{\mu}_i - \boldsymbol{\mu}_{i-1}||}{\Delta t_i}, \frac{\sqrt{2}\sigma_l}{\Delta t_i}\right).$$

Two example Rician speed distributions are shown in Figure 4. When the Rician's mean is sufficiently larger than its standard deviation, the Rician can be approximated by a Gaussian with the same mean and standard deviation. Thus, we model the measurement noise as a Gaussian with standard deviation $\sigma_{s_{measurement}} = \frac{\sqrt{2}\sigma_l}{\Delta t_i} = 2.12$ m/s $= 4.74$ mph.

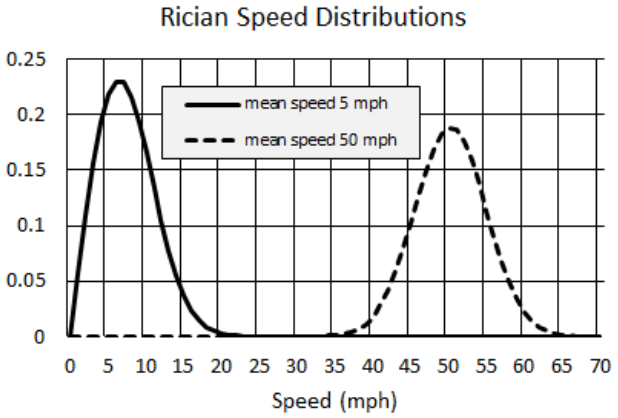

Figure 4: These are two Rician speed distributions $P_{s_i}(s_i)$, assuming GPS precision of σ_g = 3 meters and a GPS sampling interval of $\Delta t_i = 2$ seconds.

With two additive sources of Gaussian noise, the variances add to give a speed report distribution of $S \sim N(\mu_s, \sigma_s^2)$, where μ_s is the mean speed from our 5-minute traffic data, and

$$\sigma_s^2 = \sigma_{s_{natural}}^2 + \sigma_{s_{measurement}}^2. \quad (3)$$

We use this value of σ_s^2 for our experimental simulations and for assigning uncertainty to speed measurements for our MRF.

5 Markov Random Field Baseline

In this section we give a baseline scenario to explore how decreasing the number of speed reports affects the accuracy of our traffic inferences. We call this a baseline, because we modify this scenario in subsequent sections for more efficiency and privacy.

5.1 Reporting Structure

We use loopy belief propagation on the MRF for estimating traffic speeds from vehicle speed reports. We test the accuracy of our inferences by simulating noisy speed reports from vehicles at a subset of all 109 measurement stations.

[1] We could also use a speed measurement from the vehicle's speedometer. However, we are assuming that a cell phone in the vehicle is used for both communication and speed measurements, eliminating the need for the phone to interface with the vehicle's speedometer.

We vary two parameters for these tests: the number of reporting stations in the subset of all measurement stations and the number of vehicles reporting from each measurement station.

We assume the central server will receive the vehicle speed reports. For those reports coming from the same station, the server will compute the mean of the reported speeds. The distribution of the mean of n reports, where each report is distributed as $N(\mu_s, \sigma_s^2)$, is $N(\mu_s, \sigma_s^2/n)$, meaning the precision is increased. Note that σ_s represents the speed uncertainty from Equation 3, due to both measurement noise and the natural variation in speeds among a group of vehicles on the same road.

5.2 Experiments

As mentioned above, we used PeMS data from the first six months of 2017 to compute the joint PDFs necessary for the MRF. This is the training phase. We used data from the last three months of 2017 for testing. Our results show the inference error over 10,000 independent tests. Each test consists of first choosing a random 5-minute reporting interval in our test data. Inside each test, we randomly shuffle the list of measurement stations and then increment from 0 to all 109 stations, gradually including speed reports from more stations. For each tested subset of stations, we compute the root mean square (RMS) speed error between the ground truth and the mode of the inferred speed PDF at each station. We also computed a demand-weighted RMS error based on the actual vehicle flow at each station. Specifically, if the speed error at a station is Δs_i, then the demand-weighted error is $\Delta s_i f_i$, where f_i is the flow in number of vehicles per five minutes. The total demand-weighted RMS error, $\sqrt{\frac{1}{N}\sum_N \Delta s_i^2 f_i^2}$, accounts for the actual number of vehicles that would experience the speed error, placing less emphasis on lightly-traveled roads and more on heavily-traveled roads.

The accuracy results are shown in Figure 7. The horizontal axis represents an increasing number of randomly shuffled reporting stations. The solid lines in both plots show how the median RMS speed error and median RMS demand-weighted speed error both fall with more stations reporting, as expected. Here the medians are taken over the 10,000 random tests. It is also apparent that increasing the number of vehicles reporting from each station reduces error. The topmost solid curve in both plots shows the error with only one vehicle reporting from each station, and the other curves show how the error decreases with $n = 5$, 10, and 20 vehicles reporting. This decrease in error is because the precision of the speed reports is reduced as σ_s^2/n. We note that demand-weighted RMS error has the awkward units of (mph)*(vehicles/(5 minutes)). For the remainder of the paper, we will abbreviate (vehicles/(5 minutes)) as simply "flow", meaning that demand-weighted RMS error has units mph*flow.

Looking more closely at error vs. the number of vehicles reporting, Figure 5 shows how speed error and demand-weighted speed error both drop with more vehicles reporting if a random subset of 54 stations (about half) report. Beyond reports from about 20 vehicles, there is not much to gain in terms of error reduction. Because of this, for the remainder of this paper, we give results for $n = 1$, 5, 10, and 20 vehicles reporting.

We can quantify the reduction in speed reports. The mean traffic flow over our 3-month test period was 240.4 vehicles/5 minutes at each measurement station. If all vehicles reported their speeds every 5 minutes at all 109 stations, the total number of reports would be $109 \times 240.4 = 26,204$ reports every 5 minutes. If only M stations report ($M \leq N = 109$) with only n vehicles per station, then the total number of reports would be Mn every five minutes. The plots in Figure 7 show that inference error drops steadily with increasing number of stations reporting, so we will have $M = 109$ to represent all stations reporting. However, the plots show that, say, $n = 20$ vehicles reporting from each station appears to be approaching the minimum error value. With these settings of M and n, the number of reports is reduced by a factor of $(109 \times 20)/(109 \times 240.4) = 0.083$. Thus the overall accuracy of the system is maintained with only about 8% of all the vehicles reporting, representing an order of magnitude reduction in privacy-compromising communication and bandwidth.

This baseline depends on some sort of vehicle-to-vehicle coordination such that only a predefined number of vehicles report their speed from every station. The next section explores a method to eliminate the necessity of communicating to coordinate.

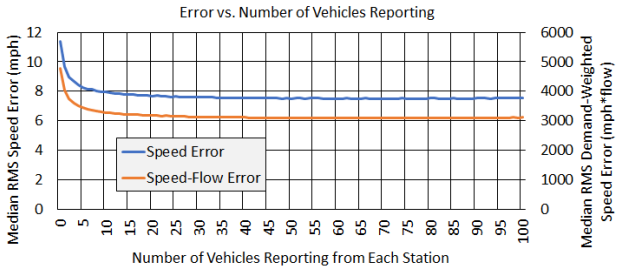

Figure 5: As more vehicles report their speeds, there are diminishing returns beyond about 20 vehicles. This plot shows computed errors for reports from random subsets of 54 stations out of all 109.

6 Coordinating Multiple Vehicle Reports Without Communication

In our scenario, vehicles each make independent decisions about whether or not to report their speed, and these decisions are renewed every five minutes. One way to reduce the number of reports is to send them only when the measured speed is unusually above or below expectations. We detail this idea in Section 7. Another way to reduce the number of reports is to have only a few vehicles on each road segment report their speed, as described in the previous section. Here we look at how to select a subset of vehicles to report without any communication from them.

6.1 Probabilistic Coordination

It is unnecessary to have all vehicles reporting. One key challenge of this scheme is how to decide which vehicles on a road segment should make a report. While vehicles could coordinate via wireless, this represents a potential privacy leak, and such coordination has yet to be standardized. Instead, we demonstrate an example of "coordination without communication" (Fenster, Kraus, and Rosenschein 1995). If a vehicle decides that a report would be useful, we have it report with a probability p_r, regardless of what other vehicles might be doing.

In a five-minute period, there are f vehicles passing a point on a road, where f indicates flow/5 minutes.[2] With f vehicles each making a report with probability p_r, the expected number of reports is $p_r f$, and the distribution of the number of vehicle reports is binomial. In particular, the probability of getting at least one report is $p_{one} = 1 - (1 - p_r)^f$, shown in Figure 6. A system operator could set p_r such that p_{one} always maintains some minimum value.

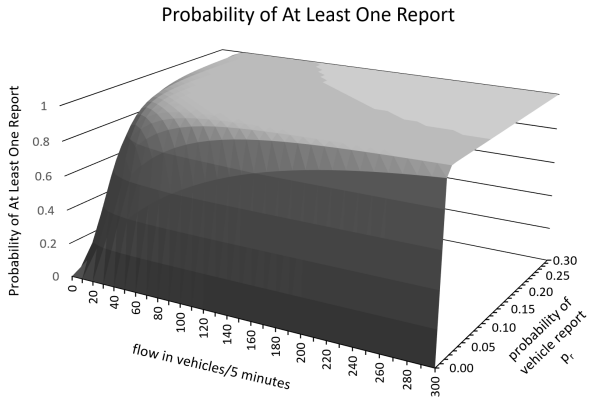

Figure 6: This is the probability of a group of vehicles sending at least one report for different group sizes and reporting probabilities.

A suitable scheme for probabilistic coordination would have a system-wide minimum threshold on the expected number of vehicle reports, $E[n]$, from every station. With $E[n] = p_r f$, each vehicle could compute its own probability of reporting as $p_r = E[n]/f$. Each vehicle would generate a uniformly distributed random number $u \in [0, 1]$ and transmit its speed to the central server if $u \leq p_r$.

The probability p_r depends on the flow f, which is not easily measurable from a moving vehicle. However, for each station i, the vehicle can estimate the flow distribution from the joint PDF $P_{s_i f_i}(s_i, f_i)$. With the vehicle measuring its own speed of $\hat{s}_i$, the flow distribution is $P_{f_i}(f_i) = P_{s_i f_i}(\hat{s}_i, f_i)$, and the scalar flow estimate is taken as the mean or mode of this distribution. We used the mode in our experiments.

[2]We could model a smaller pool of potential vehicle reports by reducing the number of potential participants to αf, with $0 < \alpha \leq 1$, but we will assume full participation with $\alpha = 1$ for our analysis.

6.2 Experiments

Our experiments for this scenario are designed to investigate the effect of multiple vehicle reports from each station and to understand the efficacy of coordinating reports probabilistically without communication among vehicles. Our baseline is the MRF model where we examine the effect of incrementally increasing the number of stations reporting, where each station has speed reports from $n = 1, 5, 10$, and 20 vehicles. This baseline is shown as the solid curves in Figure 7.

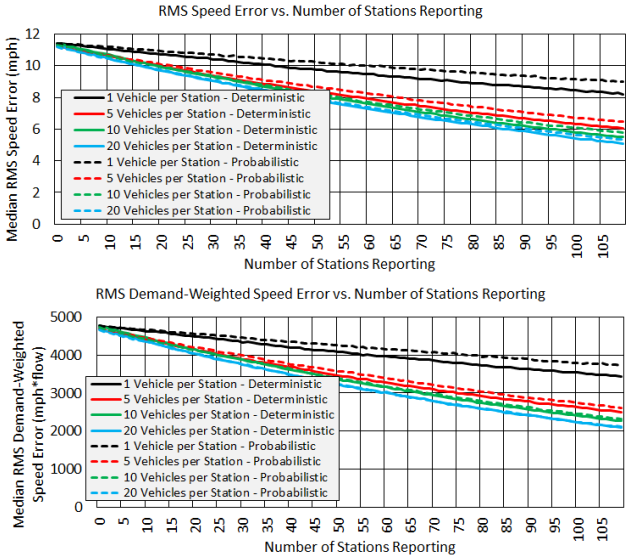

Figure 7: The solid lines show how the speed error (top plot) and demand-weighted speed error (bottom plot) change with the number of stations reporting and the number of vehicles reporting from each station. The corresponding dashed lines show the same errors when all the vehicles coordinate probabilistically without any communication from them.

Using the probabilistic reporting technique from above, Figure 7 shows that it works almost as well as its deterministic counterpart. In this figure, the dashed lines represent the probabilistic technique, and their colors colors correspond to the deterministic technique with the same target number of vehicles reporting.

Summarizing, Section 5 shows that only a relatively small subset of vehicles need to report from each station, and this section shows that the subset can be self-selected with no need for communication, enhancing privacy and reducing bandwidth. Next we describe a technique to prioritize which stations should report, which can further decrease the number of speed reports while still maintaining accuracy.

7 Prioritizing Reports with Value of Information

Not all information is equally valuable for inference tasks. In our case, some station reports are more important than others for inferring accurate, network-wide traffic conditions. Intuitively, if traffic is moving as expected, then there is little need to tell anyone. We aim to have each vehicle assess

the potential value of reporting its own speed. Each vehicle would do this locally, without any transmissions. Then, in cooperation with the central server, vehicles would either transmit their speed reports or not, based on their local self-assessments. When some vehicles choose not to report, this further reduces the communication necessary for maintaining accurate traffic inferences.

7.1 Defining Value of Information

We define the value of information (VOI) of a vehicle speed report as the RMS of the reduction in demand-weighted speed error over the road network. To define this precisely, we begin by introducing a function $s'_i(S)$ which represents the MRF's estimate of the speed at station i based on a set of speed reports S from a subset of stations in the network. The expression $s'_i(\varnothing)$ is the MRF estimate of of s_i based on no speed reports.

Ideally we would compute a quantity like Equation 4, which reflects the RMS error between the nominally expected speeds, $s'_j(\varnothing)$, and the actual speeds s_j, weighted by the actual flows f_j:

$$VOI = \sqrt{\frac{1}{N}\sum_{j=1}^{N} f_j^2 (s'_j(\varnothing) - s_j)^2} \quad (4)$$

This quantity gives the error between the ground truth speeds and the nominally expected speeds from the MRF if there were no speed reports, assuming ground truth speeds and flows are available. However, an independently operating vehicle does not have access to the actual speeds and flows, so it must estimate VOI based on what it can measure.

7.2 Estimating Value of Information

Our algorithm depends on each vehicle being able to assess the VOI of its own speed report. Because the VOI from Equation 4 depends on ground truth values of speed and flow (s_j and f_j) over the whole network, individual vehicles could not compute VOI independently. However, they can estimate VOI by inferring the speeds and flows over the network using an MRF. Specifically, a vehicle at station i can use its own measured speed $\hat{s}_i$ to compute an estimate of any speed and flow in the network with $s'_j(\{\hat{s}_i\})$ and $f'_j(\{\hat{s}_i\})$, respectively. Using these self-estimated values of speed and flow, the vehicle can then estimate the network-wide VOI of its own speed report as

$$VOI'(\hat{s}_i) = \sqrt{\frac{1}{N}\sum_{j=1}^{N} (f'_j(\{\hat{s}_i\})^2 (s'_j(\varnothing) - s'_j(\{\hat{s}_i\})^2} \quad (5)$$

$VOI'(\hat{s}_i)$ can be considered as an estimate of the amount of surprise that would be caused by transmitting speed report $\hat{s}_i$ to the central server. If the estimated VOI is high, then the system would be be surprised by the report compared to its nominal traffic conditions.

The system-wide VOI estimate in Equation 5 is perched tenuously on a single speed measurement. In order to boost accuracy, we introduce a machine-learned regression model that uses two numerical features to make a revised estimate of VOI, called VOI''. The first of these features is the original estimated $VOI'(\hat{s}_i)$. The second feature is the elapsed time from the start of the most recent Monday, called Δt_M. This helps contextualize the original VOI estimate in time, accounting for the possibility that traffic surprises may be more or less important at certain times of the week. Thus we have a final VOI estimate as $VOI''(\hat{s}_i, \Delta t_M)$. We implemented this regression as a forest of boosted decision trees, and we learned one model for each of the 109 measurement stations using months 7-9 of our PeMS data for training. Table **??** shows the important parameters that we used for this forest of trees. The median absolute VOI prediction error over all 109 stations was 0.058 mph*flow, based on an 80/20 train/test split of the PeMS data from months 7-9. Figure 8 shows a plot of the ground truth and estimated VOI from our models.

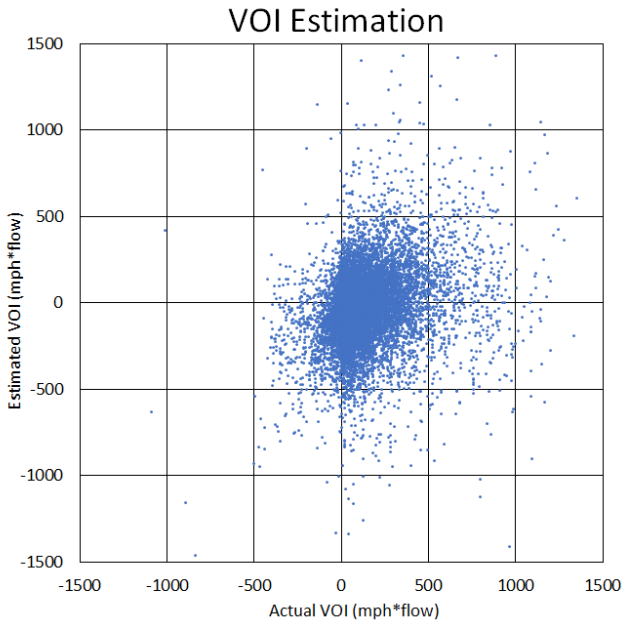

Figure 8: These are the actual and estimated VOI values from our learned regression model. The median absolute error was 0.058 mph*flow.

7.3 Experiments

One way to use the VOI estimates is to eliminate speed reports that fall below a preset VOI threshold. We computed the accuracy rate for the inferred VOI values $VOI''(\hat{s}_i, \Delta t_M)$. An accurate inference is when the inferred VOI says an instance is below a given VOI threshold and the VOI actually is below the threshold. Figure 9 shows this test accuracy rate using the same 80/20 train/test split as above. The rate is consistently above 0.965. This plot also shows that over 90% of the actual VOI values are below 1.0 mph*flow. This means that a large majority of speed re-

ports have little value and that our VOI inference procedure can accurately identify these useless reports.

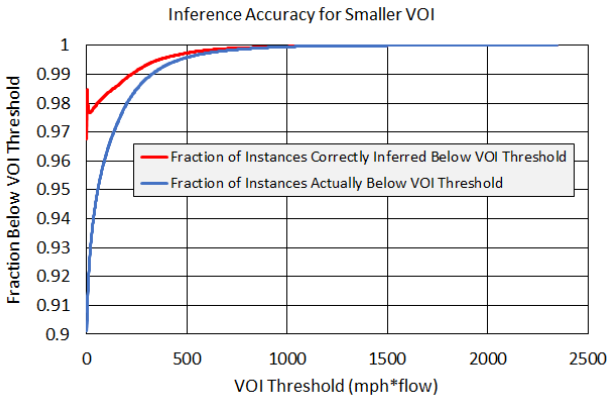

Figure 9: The VOI inference is accurate at finding speed reports that are below a given VOI threshold (red curve). Also, a large majority of potential speed reports have a very low VOI (blue curve). The VOI inferences can accurately identify these useless reports. Note that the vertical axis starts at 0.90.

We also tested our VOI approach with speed reports using months 10-12 of our traffic data, which is the same test data we used for testing in Section 6. For each randomly chosen date and time in the test set, we first incremented through a randomly shuffled list of stations. This baseline shows how the RMS speed error and demand-weighted speed error decrease as more stations report. To test the VOI approach, we used $VOI''(\hat{s}_i, \Delta t_M)$ to estimate the VOI of a report from each station. The stations then reported in descending order of estimated VOI, which was designed to test the effect of processing the most important reports earlier. In a real system, we would implement this greedy approach in a way that does not require vehicles to transmit anything until their report was needed. For instance, the central server might send out a decreasing sequence of VOI thresholds, and vehicles would respond with a report when their estimated VOI exceeded the broadcast threshold.

The results of these tests are shown in Figure 10. The VOI approach gives consistently lower RMS errors than the random approach, validating the effectiveness of this method. Although the improvement seems small in Figure 10, looking at the results another way shows a relatively dramatic improvement. In Figure 11, the horizontal axis is the number of station reports from the VOI method. The vertical axis shows how many more reports the random approach would need to equal the RMS error of the VOI approach. The number of station reports saved by the VOI approach reaches as high as 23 and has an average value of about 17.4 when there is one vehicle reporting from the stations.

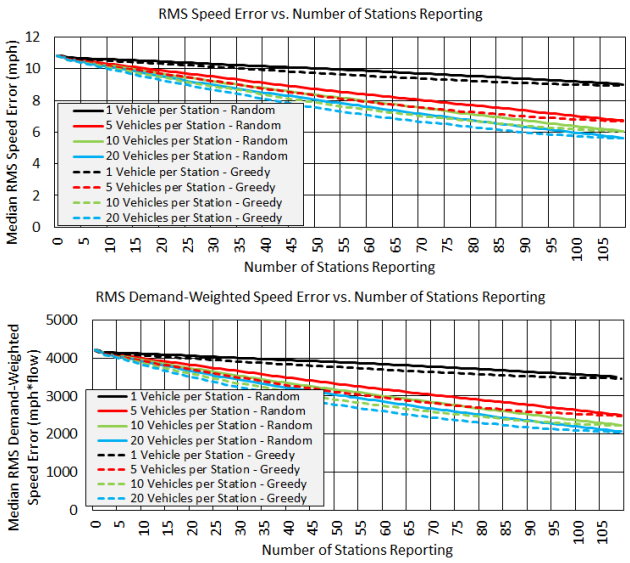

Figure 10: The solid lines show how the speed error (top plot) and demand-weighted speed error (bottom plot) change with the number of randomly-ordered stations reporting and the number of vehicles reporting from each station. The corresponding dashed lines show the corresponding errors when the stations report in descending order of estimated VOI. The dashed lines always show lower error than their solid counterparts, meaning the VOI method gives consistently lower error.

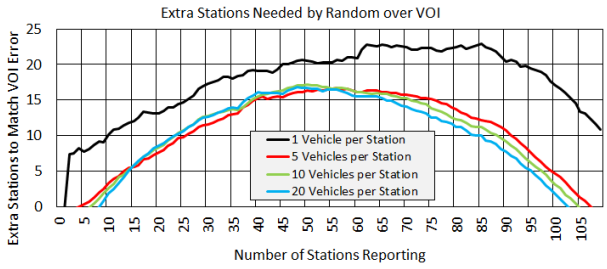

Figure 11: To match the same error as the VOI approach, this plot shows how many extra randomly chosen measurement stations would be necessary.

8 Conclusion

Naively soliciting speed reports from all eligible vehicles on the road reduces privacy and increases bandwidth requirements. This paper shows how to reduce the number of speed reports while still maintaining accurate traffic inferences. We developed and tested a Markov random field to model traffic in terms of speed and flow. The MRF has the flexibility to process speed reports from an arbitrary subset of measurement stations, with each measurement having arbitrary uncertainty. Using this model, we showed that a relatively small number of vehicles need to report from each measurement station, with about 20 vehicles per station nearing the

point of diminishing returns. We also showed how to coordinate which subset of vehicles transmit a report without requiring any explicit coordination or communication among them, leading to only a slight decrease in inference accuracy and a boost in privacy. Another method to decrease the number of reports is to estimate the value of information of each report before transmitting. Our VOI estimation algorithm can run using only a single vehicle's own speed measurement, using the MRF to infer how the report will affect the traffic inference over the whole road network. Predicting a report's value with a machine-learned regression model, we can prioritize speed reports and gain an accuracy advantage over choosing reports at random. As part of our investigation into VOI, we found that about 90% of potential speed reports in our test set were useless and that our VOI inference method can correctly detect these useless reports with high accuracy. Using VOI to prioritize reports, we can reduce the number of required reports to achieve the same error level as a randomly chosen subset of reports.

Future work along these directions could include methods for constructing an MRF using a subset of edges that balances inference accuracy (more edges) and inference speed (fewer edges). While our MRF covers only a part of the entire road network, a more general version of this method may cover a larger set of roads, possibly with overlapping MRFs. Finally, it would be instructional to investigate specific traffic anomalies, such as vehicle emergencies, to understand how a system like ours could respond dynamically to sudden changes in information needs and surprise.

References

Akcelik, R. 1996. Relating flow, density, speed and travel time models for uninterrupted and interrupted traffic. *Traffic Engineering+ Control* 37(9):511–16.

Center, P. R. 2018. Mobile fact sheet. `http://www.pewinternet.org/fact-sheet/mobile/`. [Online; accessed 1-June-2018].

Chen, P.-T.; Chen, F.; and Qian, Z. 2014. Road traffic congestion monitoring in social media with hinge-loss markov random fields. In *Data Mining (ICDM), 2014 IEEE International Conference on*, 80–89. IEEE.

Diggelen, F. V. 2007. System design & test-gnss accuracy-lies, damn lies, and statistics-this update to a seminal article first published here in 1998 explains how statistical methods can create many different. *GPS world* 18(1):26–33.

Fenster, M.; Kraus, S.; and Rosenschein, J. S. 1995. Coordination without communication: Experimental validation of focal point techniques. In *ICMAS*, 102–108.

Horvitz, E. J.; 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*.

Hu, H.; Li, G.; Bao, Z.; Cui, Y.; and Feng, J. 2016. Crowdsourcing-based real-time urban traffic speed estimation: From trends to speeds. In *Data Engineering (ICDE), 2016 IEEE 32nd International Conference on*, 883–894. IEEE.

Iannone, G.; Guarnaccia, C.; and Quartieri, J. 2013. Speed distribution influence in road traffic noise prediction. *Environmental Engineering And Management Journal* 12(3):493–501.

Kamijo, S.; Matsushita, Y.; Ikeuchi, K.; and Sakauchi, M. 2000. Traffic monitoring and accident detection at intersections. *IEEE transactions on Intelligent transportation systems* 1(2):108–118.

Kataoka, S.; Yasuda, M.; Furtlehner, C.; and Tanaka, K. 2014. Traffic data reconstruction based on markov random field modeling. *Inverse Problems* 30(2):025003.

Kockelman, K. K., and Ma, J. 2010. Freeway speeds and speed variations preceding crashes, within and across lanes. In *Journal of the Transportation Research Forum*, volume 46.

Krumm, J. 2009. A survey of computational location privacy. *Personal and Ubiquitous Computing* 13(6):391–399.

Lelewer, D. A., and Hirschberg, D. S. 1987. Data compression. *ACM Computing Surveys (CSUR)* 19(3):261–296.

Liu, Z.; Li, Z.; Li, M.; Xing, W.; and Lu, D. 2016. Mining road network correlation for traffic estimation via compressive sensing. *IEEE Transactions on Intelligent Transportation Systems* 17(7):1880–1893.

Nagatani, T. 2002. The physics of traffic jams. *Reports on progress in physics* 65(9):1331.

of Transportation (Caltrans), C. D. 2018. Performance measurement system (pems). `http://pems.dot.ca.gov/`. [Online; accessed 26-April-2018].

Parmy Olson, F. M. 2014. Why google's waze is trading user data with local governments. *Forbes Magazine*. [Online; accessed 1-June-2018].

Rice, S. O. 1945. Mathematical analysis of random noise. *The Bell System Technical Journal* 24(1):46–156.

Waze. 2016. Privacy and waze. `https://blog.waze.com/2016/04/privacy-and-waze.html`. [Online; accessed 1-June-2018].

Williams, J. 2017. Danger ahead: The government's plan for vehicle-to-vehicle communication threatens privacy, security, and common sense. `https://www.eff.org/deeplinks/2017/05/danger-ahead-governments-plan-vehicle-vehicle-cor` [Online; accessed 1-June-2018].

Yedidia, J. S.; Freeman, W. T.; and Weiss, Y. 2003. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium* 8:236–239.

Zhang, J.; Zheng, Y.; and Qi, D. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*, 1655–1661.

Zhu, H.; Zhu, Y.; Li, M.; and Ni, L. M. 2009. Seer: Metropolitan-scale traffic perception based on lossy sensory data. In *INFOCOM 2009, IEEE*, 217–225. IEEE.

Zhu, Y.; Li, Z.; Zhu, H.; Li, M.; and Zhang, Q. 2013. A compressive sensing approach to urban traffic estimation with probe vehicles. *IEEE Transactions on Mobile Computing* 12(11):2289–2302.