Real-Time Privacy-Preserving Model-Based Estimation of Traffic Flows

Jerome Le Ny *
Department of Electrical Engineering & GERAD
Polytechnique Montréal
Montréal, QC H3T 1J4, Canada
jerome.le-ny@polymtl.ca

Ahmed Touati
Department of Mathematics
ENS Cachan
61, Av du President Wilson
94235 Cachan, France
ahmed.touati@ens-cachan.fr

George J. Pappas
Department of Electrical and Systems Engineering
University of Pennsylvania
200 South 33rd Street
Philadelphia, PA 19104, USA
pappasg@seas.upenn.edu

ABSTRACT

Road traffic information systems rely on data streams provided by various sensors, e.g., loop detectors, cameras, or GPS, containing potentially sensitive location information about private users. This paper presents an approach to enhance real-time traffic state estimators using fixed sensors with a privacy-preserving scheme providing formal guarantees to the individuals traveling on the road network. Namely, our system implements *differential privacy*, a strong notion of privacy that protects users against adversaries with arbitrary side information. In contrast to previous privacy-preserving schemes for trajectory data and location-based services, our procedure relies heavily on a macroscopic hydrodynamic model of the aggregated traffic in order to limit the impact on estimation performance of the privacy-preserving mechanism. The practicality of the approach is illustrated with a differentially private reconstruction of a day of traffic on a section of I-880 North in California from raw single-loop detector data.

Categories and Subject Descriptors

C.3 [**Special-Purpose and Application-Based Systems**]: Signal processing systems; G.3 [**Probability and Statistics**]: Time series analysis

*Corresponding author. This work was done while the second author was visiting Polytechnique Montreal.

ICCPS'14, April 14–17, 2014, Berlin, Germany.

General Terms

Security, Algorithms

Keywords

Differential privacy, intelligent transportation systems, privacy-preserving data assimilation

1. INTRODUCTION

Traffic congestion remains one of the major concerns in urban areas around the world. This phenomenon is responsible for large and increasing costs linked to lost time, fuel consumption, or increased pollution and health-related issues. Considering the difficulty of developing new expensive infrastructures, it is crucial to develop better ways of managing traffic, e.g., using real-time control policies. These strategies can be enabled by the proliferation of sensors available to provide a more accurate picture of the traffic state over time. Such sensors include traditional systems built into the pavement, e.g., single and double loop inductors, more recent systems such as cameras or radars integrated into the infrastructure, as well as floating cars moving with the traffic and periodically reporting their position and speed information. Recently, there has also been some interest in letting individual users with GPS equipped smartphones play the role of such floating cars, sending their personal location traces to a data aggregator responsible for the traffic state estimation [7].

One issue that is generally not formally addressed in the vast literature on traffic state estimation is that of privacy. However, as with other monitoring systems relying on increasingly detailed data from their participants, such as smart grids and various sensor networks, privacy related concerns are starting to be voiced, and should be carefully addressed in order to encourage user adoption of these systems. With traffic estimation systems or location based services, the main concern is related to the potential possibility for an adversary to ob-

tain information about or even reconstruct a given individual's trajectory from publicly available data. This data includes what the aggregator or service provider decides to publish, e.g., real-time traffic density maps, but also all accessible databases containing information that could be linked for inference purposes. For example, the knowledge of home and work location of individuals can easily help re-identify anonymized location traces [24]. The great diversity of the sources of side information renders the task of of modeling privacy attacks very challenging.

Previous work on preserving individuals' privacy in traffic information systems is sparse. Hoh et al. [8] rely on a notion of privacy, k-anonymity, that is not particularly strong at preserving location privacy [17]. In particular, they focus on privacy for individual measurements, and thus do not directly offer formal protection guarantees for users transmitting time-series such as location traces. Some research on privacy for location-based services, see e.g., [2, 14], can be considered somewhat related to our work. These papers are typically concerned with perturbing GPS location traces to provide privacy while reconstructing some aggregate statistics, e.g, average density. However, they generally either do not rely on a formal definition of privacy, are not well adapted to real-time computations, or consider simply the minimization of mutual information between the users' private data and the published data, which ignores the crucial issue of side information [16]. Note also that most recent emphasis has been on privacy issues raised by GPS-equipped smartphones, whereas we only consider here more traditional static sensors. However, the latter can a priori also pose a risk, especially as their density increases. Indeed, [1] describes an optimization-based vehicle reidentification scheme that uses all types of traffic measurements.

The goal of this paper is to enhance a traffic state estimator, producing in real-time a traffic density or speed profile for a road section, with a privacy-preserving scheme providing quantitative privacy guarantees. For this purpose, we rely on a formal and strong notion of privacy, called *differential privacy*, which has recently been widely adopted for many applications involving sensitive personal data [4,6]. This notion of privacy characterizes certain randomized algorithms that produce answers to statistical queries according to a distribution that is not very sensitive to the presence or absence of the data of any single individual. As a result, people contributing to a dataset are guaranteed that by doing so, they do not dramatically increase the ability of an adversary to infer private information about them, even by linking the released answers to any other available sources of information. Moreover, compared to previous research on privacy for location based services, a distinguishing feature of our work is that we do not only rely on the microscopic data originating from the participants, but also on a macroscopic model of the aggregate dynamics, which helps reduce the degradation in estimation performance due to the privacy-preserving mechanism. While this idea is standard for traffic state estimation, model-based methods for differential private data assimilation are still underdeveloped.

The rest of this paper is organized as follows. Sections 2 and 3 present some background material necessary to develop a model-based traffic state estimator. Namely, we introduce a macroscopic model of traffic flow dynamics, and the measurement models for the data originating from single-loop detectors, the only type of sensors considered here. Section 4 also provides some background on designing differentially private mechanisms. Section 5 explains how we sanitize the sensor measurements to provide differential privacy guarantees. Finally, Section 6 describes the overall privacy-preserving estimator architecture, built around an extended Kalman filter and a traffic mode estimator. We also present some results on estimating traffic from real world induction loop data on a section of the Interstate 880 in California. Note that while we assume a particular traffic estimator for concreteness, the privacy-preserving scheme should be applicable to other choices.

Notation: Throughout the paper, $\mathbb{P}$ denotes a probability measure defined on a generic probability triple $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-algebra on Ω. $[N]$ denotes the set $\{1, \dots, N\}$.

2. TRAFFIC FLOW DYNAMICS

In this section we present the model of the traffic flow dynamics that we use in our data-assimilation scheme. The reader can refer to [20] for example for a detailed discussion of such models. We consider unidirectional traffic along a single road section, with position denoted x. We can have a possibly varying number of lanes $\lambda(x)$, and neglect the influence of on- and off-ramps for simplicity. Denoting by $\rho_{\text{tot}}(x,t)$ the total density over all lanes (in vehicles per mile say) and by $q_{\text{tot}}(x,t)$ the total flow over all lanes, we have the conservation law for vehicles, in integral form,

$$\frac{d}{dt}\int_{x_1}^{x_2} \rho_{\text{tot}}(x,t)dx = q_{\text{tot}}(x_1,t) - q_{\text{tot}}(x_2,t), \quad \forall x_1, x_2, t, \tag{1}$$

and the corresponding partial-differential equation (PDE) form valid at points where the solutions are sufficiently smooth

$$\frac{\partial \rho_{\text{tot}}}{\partial t} + \frac{\partial q_{\text{tot}}}{\partial x} = 0. \tag{2}$$

For numerical simulations purposes, finite-difference methods approximating (2) are generally not appropri-

ate due to the existence of discontinuities (shock waves) in the solutions. Among various discretization methods available for conservation laws, we have finite volume methods that divide the road into cells, and compute the average density in each cell recursively. We thus divide the road into I cells numbered $1, \dots I$. We add two so-called *ghost cells* numbered 0 and $I+1$, one on each side, to handle boundary conditions, which will be discussed later. The discrete-time conservation law for vehicles corresponding to the integral form (1), is then

$$\rho^i_{\text{tot},k+1} = \rho^i_{\text{tot},k} + \frac{\tau}{L_i}(f^{i-1}_{\text{tot},k} - f^i_{\text{tot},k}), \quad \text{for } i = 1, \dots I, \tag{3}$$

where τ is a timestep, L_i is the length of cell i, $f^i_{\text{tot},k}$ is the total so-called *numerical flux* out of cell i (i.e., through the interface $i \to i+1$) during period k, and $\rho^i_{\text{tot},k}$ is the total vehicle density in cell i at period k, i.e., during the time interval $[k\tau, (k+1)\tau)$. Note here that the numerical flux $f^i_{\text{tot},k}$ is different in general from the flow $q_{\text{tot}}(x_{i|i+1}, t)$, where $x_{i|i+1}$ denotes the location of the interface between cells i and $i+1$. More details are provided below.

To complete the model, we then need to introduce a hypothesis on driving behavior, typically expressed in the form of a relation between flow or speed and density. However, these relations must normally be expressed in terms of lane-averaged, also called *effective*, quantities. Hence we define the lane-averaged traffic density $\rho(x,t)$ (say, in vehicles per mile per lane), lane-averaged traffic speed $v(x,t)$, and lane-averaged traffic flow $q(x,t) = \rho(x,t)v(x,t)$ [20, Chapter 7]. If we denote by $\rho_j(x,t)$, $v_j(x,t)$ and $q_j(x,t)$ the density, speed and flow in lane j at a position x, and recalling that $\lambda(x)$ is the number of lanes, then we have the relations

$$\rho = \frac{\sum_{j=1}^{\lambda} \rho_j}{\lambda} =: \frac{\rho_{\text{tot}}}{\lambda}, q = \frac{\sum_{j=1}^{\lambda} q_j}{\lambda}, v = \sum_{j=1}^{\lambda} \frac{\rho_j}{\rho_{\text{tot}}} v_j.$$

The discretization (3) remains valid with the effective (lane-averaged) density and flux ρ_k and f_k replacing the total quantities, except in regions where the number of lanes changes. For the discrete model, we define λ^i to be the number of lanes at the interface $i \to i+1$. Any location where the number of lanes changes is always assumed to fall inside a cell. The modified discrete model for effective quantities with a varying number of lanes is then [20, p. 74]

$$\rho^i_{k+1} = \rho^i_k + \frac{\tau}{L_i}\left(\frac{\lambda^{i-1}}{\lambda^i} f^{i-1}_k - f^i_k\right), \quad \text{for } i = 1, \dots I. \tag{4}$$

We can now provide an expression of the effective numerical flux f^i_k through the interfaces. In first-order models, proposed initially by Lighthill and Whitham [11] and independently by Richards [15] (LWR models),

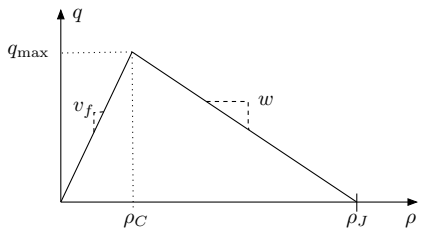

Figure 1: Triangular fundamental diagram and associated parameters.

the effective density is a fundamental quantity and a sufficient description of the local traffic state, since the effective speed and thus also the effective flow are assumed to be known static functions of it. The assumed expression of flow in terms of density $q(\rho)$ is called the *fundamental diagram*, and can be obtained for a specific road segment for example by fitting observational data.

In this paper we work for concreteness with triangular fundamental diagrams, which are arguably the most popular in practice. The resulting LWR model is also called the *Cell-Transmission Model* (CTM) [3], and can be efficiently simulated. For simplicity we make the additional modeling assumption that each cell has the same fundamental diagram, whose parameters are assumed known. The triangular fundamental diagram can be expressed as follows

$$q(\rho) = \begin{cases} v_f \rho & \text{if } \rho \le \rho_C = \frac{w}{v_f + w}\rho_J, \\ w(\rho_J - \rho) & \text{if } \rho_C < \rho \le \rho_J, \end{cases}$$

where ρ_J is the maximum or "jam" density on the road segment, and ρ_C is the critical density at which the maximum flow $q_{\max} = v_f \rho_C$ is attained. The two cases correspond to free and congested traffic respectively, with v_f the free traffic speed, and w the congestion wave propagation speed. Fig. 1 illustrates these definitions.

For this triangular fundamental diagram, a standard numerical method, the Godunov method, corresponds to using the following numerical flux in (4):

$$\begin{aligned} f^i_k = F(\rho^i_k, \rho^{i+1}_k) &:= \min\{S(\rho^i_k), R(\rho^{i+1}_k)\} \\ &= \min\left\{\rho^i_k v_f, q_{\max}, w(\rho_J - \rho^{i+1}_k)\right\}. \end{aligned} \tag{5}$$

Note that this flux can be interpreted as the minimum between the maximum flow $S(\rho) = \min\{\rho v_f, q_{\max}\}$ that can be sent from cell $i-1$ and maximum flow $R(\rho) = \min\{q_{\max}, w(\rho_J - \rho^{i+1}_k)\}$ that can be received by cell i.

Starting from the deterministic CTM model, we now form the following *stochastic* state-space model of the

density dynamics on the road

$$\rho_{k+1}^i = \rho_k^i + \frac{\tau}{L_i}\left(\frac{\lambda^{i-1}}{\lambda^i}F(\rho_k^{i-1},\rho_k^i) - F(\rho_k^i,\rho_k^{i+1})\right) + \xi_k^i, \tag{6}$$

for $i = 1,\dots I$. Here ξ_k^i is a Gaussian random variable, whose variance can be tuned later on in the design of the state estimator, based on the relative confidence we place in the model or the observations. To close the model, we also need the densities ρ^0 and ρ^{I+1} in the ghost cells upstream and downstream of the road section. It is often assumed in the literature that these densities are exactly known, but this can be hard to justify in practice, and density pseudo-measurements can be very noisy. Hence we assume a simple random walk model for these densities

$$\rho_{k+1}^0 = \rho_k^0 + \xi_k^0, \;\; \rho_{k+1}^{I+1} = \rho_k^{I+1} + \xi_k^{I+1}, \tag{7}$$

where ξ_k^0, ξ_k^{I+1} denote again Gaussian random variables. This standard modeling technique allows us to estimate these ghost cell densities from the measurements as well, and is used in the context of traffic estimation for example in [21].

3. SENSORS AND MEASUREMENT MODELS

In this section, we describe the measurements available from single loop detectors, which are the most common types of sensors available for traffic estimation purposes. We then present measurement models for these sensors, which are necessary for the development of our estimator. Single loop detectors are static sensors, placed at a fixed location along the road segment and continuously observing the traffic at that location. They report *for each lane* the following quantities at periodic intervals, e.g., every $T = 30s$ or $60s$:

- Vehicle counts $c(t)$, that is, the number of vehicles that crossed the sensor line during the last period of T seconds;
- Occupancy $o(t)$, that is, the percentage of the period in the last T s where a vehicle was driving above the detector.

Single loop detectors cannot measure traffic density or velocity at their location directly. However, we can use their measurements to obtain estimates of these quantities, subject to some random errors. Namely, we have for each lane

$$v_j(t) \approx g\frac{c_j(t)}{o_j(t)T}, \;\; q_j(t) \approx \frac{c_j(t)}{T}, \;\; \rho_j(t) \approx \frac{o_j(t)}{g}. \tag{8}$$

Note that the speed and density "pseudo-measurements" depend on the so-called g *factor*, which is the average effective vehicle length at the sensor location, can vary with time, but is sometimes estimated precisely and available as public information [9], notably in the California highway system.

In our model, we include a cell boundary at each static sensor location. We then consider the pseudo-measurements (8) as the available measurements at that location, from which we want to build our effective density estimate along the road. There are therefore various ways of exploiting the loop detector measurements. The simplest measurement model is

$$y_k^i = y_k^{i+1} := \frac{1}{g\lambda^i}\sum_{j=1}^{\lambda^i} o_{j,k}^i = \rho_k^i + \nu_k^i = \rho_k^{i+1} + \nu_k^{i+1}, \tag{9}$$

where ν_k^i, ν_k^{i+1} are Gaussian random variables describing measurement errors, and y_k^i, y_k^{i+1} play the role of the density measurements for cells i and $i+1$ around the sensor placed at the interface $i \to i+1$. Note that for simplicity we choose τ to be a divisor of T. During the interval between sensor reports, we receive no measurement and thus only use the dynamic model to update the density estimate.

A second measurement model that we can develop with single-loop detector data uses the flow pseudo-measurements obtained via the vehicle counts as well as the fundamental diagram. For a sensor again at the interface $i \to i+1$, we could consider the nonlinear measurement model

$$\phi_k^i := \frac{1}{T\lambda^i}\sum_{j=1}^{\lambda^i} c_{j,k}^i = F(\rho_k^i, \rho_k^{i+1}) + \nu_k^i. \tag{10}$$

However, this model requires comparing the densities ρ_k^i and ρ_k^{i+1} or their estimates, in order to choose the correct minimum term in the expression (5), and an error in this choice can have a large impact on estimation performance.

Another approximate measurement model turned out to be more robust. Define ϕ_k^i as in (10), and assume first that we have a traffic mode estimate m_k^i for the interface, either free (F, for $\rho \le \rho_c$) or congested (C, for $\rho > \rho_c$). This estimate could be obtained by relying on the measurements y_k^i of (9) and comparing them to ρ_c, with some additional filtering described in Section 6. We then form the new density pseudo-measurements

$$z_k^i = z_k^{i+1} := \begin{cases} \frac{\phi_k^i}{v_f}, & \text{if } m_k^i = \mathsf{F} \\ \rho_J - \frac{\phi_k^i}{v_f}, & \text{if } m_k^i = \mathsf{C}, \end{cases} \tag{11}$$

and consider again the model

$$z_k^i = z_k^{i+1} = \rho_k^i + \eta_k^i = \rho_k^{i+1} + \eta_k^{i+1}, \tag{12}$$

where η_k^i, η_k^{i+1} are Gaussian random variables.

A last measurement model could use the speed pseudo-measurements involving the ratio of counts and occu-

pancies. Such a model is left for future work however. As we discuss in Section 5, occupancy measurements appear to be more difficult to handle from a differential privacy point of view, and both (9) and the velocity estimates rely directly on them. Note also that double loop detectors can measure vehicle speed directly but unfortunately they are not as common as single loop detectors. Such sensors could be advantageous from a location privacy perspective.

4. DIFFERENTIAL PRIVACY

In this section we review the notion of differential privacy [6], and present certain mechanisms that can be used to achieve it. We refer the reader to the surveys by Dwork, e.g., [4], for a more detailed discussion of the notion of differential privacy, and to [10] for the proofs of some additional results presented in Sections 4.1 and 4.2.

4.1 Definition

Let D be a space of datasets of interest (e.g., a space of data tables, or a signal space). A *mechanism* is a map $M : \mathsf{D} \times \Omega \to \mathsf{R}$, for some measurable output space R, such that for any element $d \in \mathsf{D}$, $M(d, \cdot)$ is a random variable, typically written simply $M(d)$. A mechanism can be viewed as a probabilistic algorithm to answer a query q, which is a map $q : \mathsf{D} \to \mathsf{R}$. Intuitively, in the following definition, D is a space of datasets of interest, and we have a symmetric binary relation Adj on D, called adjacency, such that $\mathrm{Adj}(d, d')$ if and only if d and d' differ by the data of a single participant.

Definition 1. Let D be a space equipped with a symmetric binary relation denoted Adj, and let $(\mathsf{R}, \mathcal{M})$ be a measurable space, where $\mathcal{M}$ is a given σ-algebra over R. Let $\epsilon, \delta \geq 0$. A mechanism $M : \mathsf{D} \times \Omega \to \mathsf{R}$ is (ϵ, δ)-differentially private (for Adj) if for all $d, d' \in \mathsf{D}$ such that $\mathrm{Adj}(d, d')$, we have

$$\mathbb{P}(M(d) \in S) \leq e^{\epsilon}\mathbb{P}(M(d') \in S) + \delta, \ \ \forall S \in \mathcal{M}. \quad (13)$$

If $\delta = 0$, the mechanism is said to be ϵ-differentially private.

The definition says that for two adjacent datasets, the distributions over the outputs of the mechanism should be close. The choice of the parameters ϵ, δ is set by the privacy policy, with smaller values corresponding to stronger privacy guarantees.

A fundamental property of the notion of differential privacy is that no additional privacy loss can occur by simply manipulating an output that is differentially private. To state this more precisely, recall that a probability kernel between two measurable spaces $(\mathsf{R}_1, \mathcal{M}_1)$ and $(\mathsf{R}_2, \mathcal{M}_2)$ is a function $k : \mathsf{R}_1 \times \mathcal{M}_2 \to [0, 1]$ such that $k(\cdot, S)$ is measurable for each $S \in \mathcal{M}_2$ and $k(r, \cdot)$ is a probability measure for each $r \in \mathsf{R}_1$. The proof of the following theorem can be found in [10].

THEOREM 1 (RESILIENCE TO POST-PROCESSING). *Let $M_1 : \mathsf{D} \times \Omega \to (\mathsf{R}_1, \mathcal{M}_1)$ be an (ϵ, δ)-differentially private mechanism. Let $M_2 : \mathsf{D} \times \Omega \to (\mathsf{R}_2, \mathcal{M}_2)$ be another mechanism, such that there exists a probability kernel $k : \mathsf{R}_1 \times \mathcal{M}_2 \to [0, 1]$ verifying*

$$\mathbb{P}(M_2(d) \in S | M_1(d)) = k(M_1(d), S), \ \ a.s., \quad (14)$$

for all $S \in \mathcal{M}_2$ and $d \in \mathsf{D}$. Then M_2 is (ϵ, δ)-differentially private.

Note that in (14), the kernel k is not allowed to depend on the dataset d. In other words, this condition says that once $M_1(d)$ is known, the distribution of $M_2(d)$ does not further depend on d. Hence a mechanism M_2 accessing a dataset only indirectly via the output of a differentially private mechanism M_1 cannot weaken the privacy guarantee. This theorem is implicitly used repeatedly in the following, every time we process an already differentially private signal to improve the quality of an estimate, while claiming that this has no impact on privacy.

Finally, the last result that we introduce relates to the independent application of several mechanisms to the same database, i.e., with each mechanism's randomness introduced independently of the others. Indeed, in our estimator, we use two differentially private mechanisms, and still want to provide a differential privacy guarantee for the overall scheme.

THEOREM 2. *Consider $M_1, \dots M_r$, r independent mechanisms on a space D, where M_i is (ϵ_i, δ_i)-differentially private. Then the mechanism $M = (M_1, \dots M_r)$, which, for $d \in \mathsf{D}$, outputs $(M_1(d), \dots, M_r(d))$, is $\left(\sum_{i=1}^r \epsilon_i, \sum_{i=1}^r \delta_i\right)$-differentially private.*

4.2 A Gaussian Mechanism for the Publication of Continuous-Valued Time Series

A mechanism that throws away all the information in a dataset is obviously private, but not useful, and in general one has to trade off privacy for utility when answering specific queries. We recall below a basic mechanism that can be used to answer numerical queries in a differentially private way. First, consider a query $q : \mathsf{D} \to \mathsf{R}$, where the output space R is equipped with a norm denoted $\|\cdot\|_{\mathsf{R}}$. The following quantity plays an important role in the design of differentially private mechanisms [6].

Definition 2. Let D be a space equipped with an adjacency relation Adj. The sensitivity of a query $q : \mathsf{D} \to \mathsf{R}$

is defined as $\Delta_{\mathsf{R}} q := \max_{d,d':\mathrm{Adj}(d,d')} \|q(d) - q(d')\|_{\mathsf{R}}$. In particular, for $\mathsf{R} = \mathbb{R}^k$ equipped with the p-norm $\|x\|_p = \left(\sum_{i=1}^k |x_i|^p\right)^{1/p}$, for $p \in [1,\infty]$, we denote the ℓ_p sensitivity by $\Delta_p q$.

A differentially private mechanism proposed in [5] modifies an answer to a numerical query by adding iid zero-mean Gaussian noise. Recall the definition of the $\mathcal{Q}$-function $\mathcal{Q}(x) := \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$. We have the following theorem [5, 10].

THEOREM 3. *Let $q : \mathsf{D} \to \mathbb{R}^p$ be a query. Then the Gaussian mechanism $M_q : \mathsf{D} \times \Omega \to \mathbb{R}^p$ defined by $M_q(d) = q(d) + w$, with $w \sim \mathcal{N}\left(0, \sigma^2 I_p\right)$, where $\sigma \geq \frac{\Delta_2 q}{2\epsilon}(K + \sqrt{K^2 + 2\epsilon})$ and $K = \mathcal{Q}^{-1}(\delta)$, is (ϵ, δ)-differentially private.*

For the rest of the paper, we define $\kappa_{\delta,\epsilon} = \frac{1}{2\epsilon}(K + \sqrt{K^2 + 2\epsilon})$, so that the standard deviation σ in Theorem 3 can be written $\sigma(\delta, \epsilon) = \kappa_{\delta,\epsilon} \Delta_2 q$. Moreover, we are interested in mechanisms publishing vector valued *signals*. Consider a system G accepting datasets and publishing such a signal based on the data, i.e., $G : \mathsf{D} \to \mathsf{R} := (\mathbb{R}^p)^{\mathbb{N}}$, for some p. The ℓ_2-sensitivity of G can be defined as in Definition 2, using the l_2 norm on R

$$\|\mathbf{x}\|_2 = \left(\sum_{k=0}^{\infty} \|x_k\|_2^2\right)^{1/2}, \text{ for } \mathbf{x} = \{x_k\}_{k\geq 0}.$$

The following theorem generalizes Theorem 3 to such systems publishing signals. Certain technical measurability issues in the proof of this result are resolved in [10]. Note that a (discrete-time) zero-mean Gaussian white noise signal w with covariance Σ is simply a sequence of independent Gaussian random variables $\{w_k\}_{k\geq 0}$ with $\mathbb{E}[w_k] = 0$, $\mathbb{E}[w_k w_l^T] = 0$ for $k \neq l$, and $\mathbb{E}[w_k w_k^T] = \Sigma$, for all k.

THEOREM 4. *Let $G : \mathsf{D} \to (\mathbb{R}^p)^{\mathbb{N}}$. The mechanism $M(d) = G(d) + n$, where n is a zero-mean Gaussian white noise with covariance matrix $\kappa_{\delta,\epsilon}^2 (\Delta_2 G)^2 I_p$, is (ϵ, δ)-differentially private.*

4.3 An Exponential Mechanism for the Publication of Discrete-Valued Sequences

Before presenting our estimator, we need to develop one more way of sanitizing data sequences, which should be applicable to discrete-valued data, such as the sequence of traffic modes over time in a cell. For this purpose we propose an extension of the exponential mechanism introduced in [12]. Consider a function $q : \mathsf{D} \times \mathsf{R} \to \mathbb{R}$, which to a dataset d and a response r associates a score $q(d, r)$. Fix a measure μ on R (we will assume μ to be uniform in this paper). The exponential mechanism, denoted $\mathcal{E}_q^{\mu,\epsilon} : \mathsf{D} \to \mathsf{R}$, given a database d, picks an output $r \in \mathsf{R}$ randomly with the probability distribution

$$\frac{\exp(\epsilon q(d, r)) d\mu(r)}{\int_{\mathsf{R}} \exp(\epsilon q(d, r)) d\mu(r)}.$$

Hence this mechanism gives a larger probability to the outputs r that maximize the score $q(d, r)$. Next, define

$$\Delta q = \sup_{r \in \mathsf{R}} \sup_{\mathrm{Adj}(d,d')} |q(d, r) - q(d', r)|.$$

We have the following theorem [12].

THEOREM 5. *The exponential mechanism $\mathcal{E}_q^{\mu,\epsilon}$ is $2\epsilon\Delta q$ - differentially private.*

PROOF. For S a measurable set of R, we have

$$\mathbb{P}(\mathcal{E}(d) \in S) = \frac{\int_S \exp(\epsilon q(d, r)) d\mu(r)}{\int_{\mathsf{R}} \exp(\epsilon q(d, r)) d\mu(r)}.$$

Now for d' adjacent to d, and for each r,

$$q(d', r) - \Delta q \leq q(d, r) \leq q(d', r) + \Delta q$$

so

$$\mathbb{P}(\mathcal{E}(d) \in S) \leq \frac{e^{\epsilon \Delta q} \int_S \exp(\epsilon q(d', r)) d\mu(r)}{e^{-\epsilon \Delta q} \int_{\mathsf{R}} \exp(\epsilon q(d', r)) d\mu(r)}$$

$$\mathbb{P}(\mathcal{E}(d) \in S) \leq e^{2\epsilon \Delta q} \mathbb{P}(\mathcal{E}(d') \in S).$$

□

The following version of the exponential mechanism will be used to estimate the traffic mode (fluid or congested) of a road segment. Consider a mechanism that publishes, for each dataset d, a sequence of vectors $\mathbf{X} := \{X_k^i\}_{k\geq 0}^{1\leq i\leq M}$, with $X_k^i \in \{\mathsf{F}, \mathsf{C}\}$ discrete-valued. Consider also a set of functions $q_k^i : \mathsf{D} \times \{\mathsf{F}, \mathsf{C}\} \to \mathbb{R}$, such that for all d, d' adjacent, there exists an increasing finite subsequence $k(i)$, $1 \leq i \leq M$, with

$$|q_{k(i)}^i(d, X_{k(i)}^i) - q_{k(i)}^i(d', X_{k(i)}^i)| \leq \rho_i, \qquad (15)$$
$$q_k^i(d, X_k^i) = q_k^i(d', X_k^i), \text{if } k \neq k(i),$$

for all $1 \leq i \leq M$. Finally, consider a mechanism $\mathcal{E}$ that, given $d \in \mathsf{D}$, picks the components X_k^i randomly and independently for all i, k, with

$$\mathbb{P}(X_k^i = \mathsf{C}) = \frac{\exp(\epsilon q_k^i(d, \mathsf{C}))}{\exp(\epsilon q_k^i(d, \mathsf{C})) + \exp(\epsilon q_k^i(d, \mathsf{F}))}. \qquad (16)$$

COROLLARY 1. *The mechanism $\mathcal{E}$ is $2\epsilon \left(\sum_{i=1}^M \rho_i\right)$ - differentially private.*

PROOF. For each finite nonnegative integer K, and each vector $\mathbf{X}_K = \{X_k^i\}_{0\leq k\leq K}^{1\leq i\leq M}$ consider the function $\hat{q}(d, \mathbf{X}_K) = \sum_{i=1}^M \sum_{k=0}^K q_k^i(d, X_k^i)$. The exponential mechanism for this function $\hat{q}$ takes exactly the equivalent

form (16). Moreover, for d, d' adjacent, there is a finite sequence $k(i)$ for $1 \leq i \leq M$ such that

$$\begin{aligned}&|\hat{q}(d, \mathbf{X}_K) - \hat{q}(d', \mathbf{X}_K)| \\ &= \left| \sum_{i=1}^{M} q^i_{k(i)}(d, X^i_{k(i)}) - q^i_{k(i)}(d', X^i_{k(i)}) \right| \\ &\leq \sum_{i=1}^{M} \rho_i.\end{aligned}$$

Hence we obtain a differentially private mechanism for each finite K. To deduce that the actual mechanism, releasing a countably infinite number of random variables, is differentially private (in other words, the case $K \to \infty$), technical measurability issues can be resolved by following the same reasoning as in [10, Lemma 2]. □

5. DIFFERENTIALLY PRIVATE TRAFFIC SENSOR MEASUREMENTS

In the rest of this paper, we consider, for datasets of N user trajectories of the form $\mathbf{x} = \{(x_1(t), \dots, x_N(t)) | t \geq 0\}$, the following adjacency relation

$$\text{Adj}(\mathbf{x}, \tilde{\mathbf{x}}) \text{ iff there exists } i \in [N] \text{ s.t. } x_j = \tilde{x}_j, \forall j \neq i. \tag{17}$$

Hence two sets of traces are adjacent according to (17) if and only if they differ by at most a single trace. Note also that here two traces differ if their value differs at any single time. Mechanisms that are differentially private for this adjacency relation are quite strong, since they essentially hide the presence of an individual, not just its location.

5.1 Flow Measurements

Consider the definition of flow pseudo-measurements ϕ^i_k in (10), based on counts reported by single-loop detectors for each lane. Let $\mathbf{x}$ and $\tilde{\mathbf{x}}$ be two sets of trajectories, adjacent for (17). Suppose that the road is equipped with M single-loop detectors, reporting at each time kT the counts $c^i_{j,k}$ or $\tilde{c}^i_{j,k}$, $i = 1, \dots M$, $j = 1, \dots \lambda^i$, corresponding to the two adjacent sets of trajectories. Denote the corresponding flow pseudo-measurements ϕ^i_k and $\tilde{\phi}^i_k$, $i = 1, \dots M$. Then

$$\|\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}\|_2^2 = \sum_{k=0}^{\infty} \sum_{i=1}^{M} |\phi^i_k - \tilde{\phi}^i_k|^2 = \sum_{i=1}^{M} \sum_{k=0}^{\infty} |\phi^i_k - \tilde{\phi}^i_k|^2.$$

Now consider say a sensor at the interfarce $i \to i+1$, and the term

$$\sum_{k=0}^{\infty} |\phi^i_k - \tilde{\phi}^i_k|^2 = \frac{1}{T^2(\lambda^i)^2} \sum_{k=0}^{\infty} \left| \sum_{j=1}^{\lambda^i} (c^i_{j,k} - \tilde{c}^i_{j,k}) \right|.$$

For concreteness assume that the differing trajectory in the two adjacent datasets is the first one, and associate to the two trajectories $x_1(t)$ and $\tilde{x}_1(t)$ two vehicles A and B. The counts $c^i_{j,k}$ and $\tilde{c}^i_{j,k}$ must be almost all identical, except for the fact that vehicles A and B can cross the line of the sensor at different periods and in different lanes. Hence $|c^i_{j,k} - \tilde{c}^i_{j,k}| = 0$ except for at most two pairs (j_A, k_A) and (j_B, k_B) where the difference can be one, corresponding to the lane and period at which A and B cross the sensor. Thus, we have

$$\sum_{k=0}^{\infty} |\phi^i_k - \tilde{\phi}^i_k|^2 \leq \frac{2}{T^2(\lambda^i)^2},$$

and finally, since the reasoning is the same for all M sensors

$$\|\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}\|_2 \leq \frac{\sqrt{2}}{T} \sqrt{\sum_{i=1}^{M} \frac{1}{(\lambda^i)^2}} =: \Delta_f. \tag{18}$$

PROPOSITION 1. *The mechanism publishing for each sensor the perturbed flow pseudo-measurements $\phi^i_k + n^i_k$, where ϕ^i_k is defined in (10) and n^i_k are independent zero-mean white Gaussian noise signals with covariance $\kappa^2_{\delta,\epsilon}\Delta_f$, with Δ_f defined in (18), is (ϵ, δ)-differentially private.*

5.2 Density and Mode Measurements

Following (8), density pseudo-measurements at the sensor location can be based on the reported occupancy measurements. Using these measurements directly is problematic however, because the occupancy time due to a single vehicle, equal to l_v/v_v, with l_v its length and v_v its speed, can vary widely depending on its speed. As a result, the sensitivity of these density pseudo-measurements is high and the standard Gaussian perturbation mechanism leads to unreliable measurements, especially at low density. This can result in frequent mode estimation errors in the final estimator.

Instead of using the occupancy measurements directly to estimate the density, the strategy adopted in this paper is to use them to estimate only the mode of the traffic, i.e., fluid or congested, which corresponds to $\rho \leq \rho_c$ and $\rho > \rho_c$ on the fundamental diagram respectively. The density itself is estimated from the flow pseudo-measurements and the fundamental diagram as in (11), but this requires an additional mode estimate as discussed here, because to each flow measurement $0 \leq q < q_{\max}$ correspond two possible densities on the fundamental diagram, namely q/v_f or $\rho_J - q/w$.

Our differentially private mode pseudo-measurements m^i_k for each sensor $i \in [M]$ at each period $k \geq 0$ are constructed using the exponential mechanism of Section 4.3. At period k and for sensor $i \in [M]$, suppose that we are given access to the raw data of occupancy measurements for each car, not just the lane aggregate (sum of occupancies) over the period. We replace each such car

measurement, say o_v, by the quantity $c_v = \min\{\frac{o_v}{g}, \rho_c\}$. Note that o_v/g is the contribution to the density pseudo-measurement due simply to car v. If this contribution exceed ρ_c, i.e., this car by itself contributes to a reading of congestion (which can happen if, say, the car passes on the sensor very slowly or stops there), we fix it at ρ_c (another value could be used, e.g., $2\rho_c$). This truncation will allow us to bound the sensitivity to individuals' data, and is not too detrimental to mode estimation, since very slowly moving cars are a good indication of congestion. Next, we replace the occupancies $o^i_{j,k}$ for each lane $1 \leq j \leq \lambda^i$, obtained by summing the o_v for the vehicles, by the quantities $c^i_{j,k}$ obtained by summing the c_v. Finally, we compute the scores

$$q^i_k(\mathsf{C}) = \frac{1}{\rho_c \lambda_i} \sum_{j=1}^{\lambda^i} c^i_{j,k}, \tag{19}$$

$$q^i_k(\mathsf{F}) = \left(2 - \frac{1}{\rho_c \lambda_i} \sum_{j=1}^{\lambda^i} c^i_{j,k}\right). \tag{20}$$

The idea is that $\frac{1}{\lambda_i}\sum_{j=1}^{\lambda^i} c^i_{j,k}$ represents an average density indicator, with the score $q^i_k(\mathsf{C})$ high for high density (congestion), and $q^i_k(\mathsf{F})$ high for low density (free traffic).

PROPOSITION 2. *The exponential mechanism publishing the mode pseudo-measurements m^i_k for each sensor, by generating them randomly and independently with identical distribution*

$$\mathbb{P}(m^i_k = \mathsf{C}) = \frac{\exp(\epsilon q^i_k(\mathsf{C}))}{\exp(\epsilon q^i_k(\mathsf{C})) + \exp(\epsilon q^i_k(\mathsf{F}))},$$

is $4\epsilon\left(\sum_{i=1}^{M} \frac{1}{\lambda^i}\right)$-differentially private.

PROOF. A single vehicle's data c_v contributes at most ρ_c to a single term $c^i_{j,k}$ in (19) or (20), since the vehicle is counted in only one lane. Changing a vehicle's trajectory can thus change the difference in (15) by at most $\frac{2}{\lambda_i}$. The assumptions of Corollary 1 are verified because a vehicle passes through each sensors' location at a unique period. □

6. TRAFFIC STATE ESTIMATION

In this last section, we present the overall architecture of our privacy-preserving traffic estimator, and illustrate its performance on a real-world dataset. The output of the estimator is a density map $\rho(x,t)$ or rather ρ^i_k, produced in real-time as time increases. From the density, we can then deduce the speed map using the fundamental diagram.

6.1 Estimator Architecture

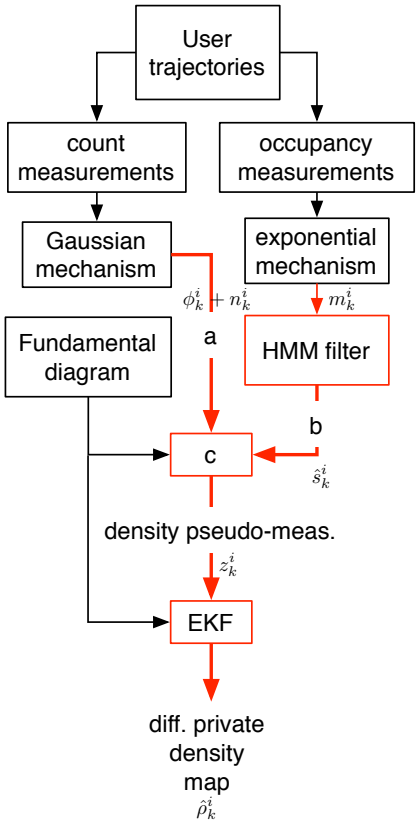

Figure 2: Architecture of our differentially private traffic estimator using single-loop detectors only. The red arrows represent differentially private signals. Legend: (a) perturbed flow pseudo-measurements from vehicle counts; (b) private mode estimate built from occupancy measurements; (c) computation (11) of density pseudo-measurements from perturbed flow measurements and mode estimates.

The estimator architecture is depicted on Fig. 2. The main data-assimilation procedure is done by an extended Kalman filter (EKF) [18], using the density pseudo-measurements z^i_k (12), and the dynamic traffic model (6). We do not use the density measurements y^i_k from (9), because a Gaussian mechanism for these measurements would require truncating the occupancy measurements as we did in Section 5.2, preventing an accurate density reading for high densities, and would still add too much noise at low densities. The differential privacy guarantee for the overall architecture is the sum of $(\epsilon_1 + \epsilon_2, \delta_1)$, where (ϵ_1, δ_1) are the parameters chosen for the Gaussian mechanism, and ϵ_2 is the parameter chosen for the exponential mechanism. Note that the architecture relies on the result of Theorem 1, which shows that the outputs of the HMM filter and the EKF are differentially private because their inputs are.

Inverting the fundamental diagram to produce density pseudo-measurements z_k^i from the (perturbed) flow measurements ϕ_k^i requires an estimate of the traffic mode, fluid of congested. This could potentially be obtained directly from the current density estimate $\hat{\rho}_k^i$ produced by the EKF, thus ignoring the occupancy measurements completely. However, this scheme was found to be unreliable, and thus the exponential mechanism described in Section 5.2 was introduced to produce mode pseudo-measurements m_k^i for each sensor from truncated occupancy measurements.

Due to the randomness introduced by the exponential mechanism however, these mode measurements m_k^i tend to switch between fluid and congested traffic too frequently. To obtain a more physically meaningful behavior, we don't use these pseudo-measurements directly but filter them through an additional hidden-Markov model (HMM) for each sensor location, described as follows. We introduce for the location of sensor i a new state trajectory $\{s_k^i\}_{k\geq 0}$, with $s_k^i \in \{\mathsf{C}, \mathsf{F}\}$, describing the actual mode estimate used to invert the fundamental diagram. The dynamics of s_k^i are simply described by a Markov chain with a single parameter $\pi_1 = \mathbb{P}(s_{k+1}^i \neq s_k^i)$ describing the probability with which the mode changes from fluid to congested at that location. This parameter could be estimated from historical data. Finally, we introduce a last parameter $\pi_2 := \mathbb{P}(m_k^i = s_k^i)$, which reflects the confidence we have in the output of the exponential mechanism, and hence should decrease from 1 to 1/2 as ϵ decreases. These parameters π_1 and π_2 describe completely the dynamics and observation model of our HMM, and are sufficient for the online recursive computation at each period of the probability distribution $\mathbb{P}(s_k^i | m_0^i, \ldots, m_k^i)$ and corresponding maximum-likelihood estimate $\hat{s}_k^i$ given the outputs $m_0^i, \ldots, m_k^i$ of the exponential mechanism.

Regarding the choice of an EKF to perform the final density estimation, we note that this filter could be replaced by one of many other nonlinear filters proposed in the literature on traffic flow estimation, see, e.g., [13, 22]. It is outside the scope of this paper to discuss particular choices of filter, although we are not aware of a study rigorously comparing the relative performance of different schemes for traffic flow estimation. The EKF is used in [21] with a different dynamic model, and in [19] with a CTM model. Yuan et al. [23] also use it with a Lagrangian model rather than the Eulerian model discussed here.

6.2 Traffic State Estimation Example

To illustrate our approach, we estimate the traffic state from induction loop data available as part of the Mobile Century experiment dataset [7]. This data consists of counts and occupancy measurements from single loop detectors, for each lane of Interstate 880 (Northbound) in California between postmile 16.5 and 27.7, i.e., along an approximately 11 mile road segment. The data from 27 detectors is available, at $T = 30s$ intervals, and the number of lanes on the road segment varies between 4 and 5.

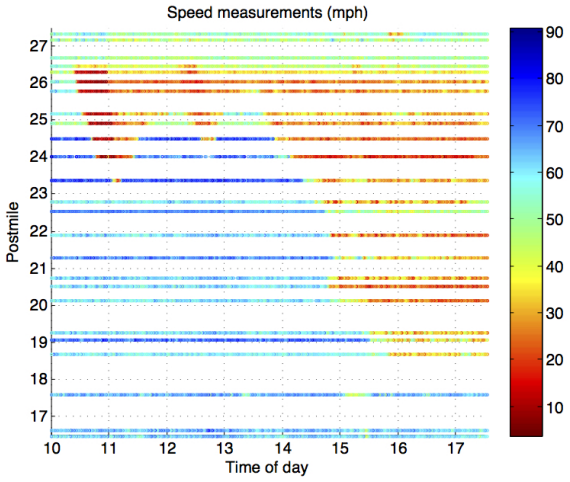

Figure 3: Speed pseudo-measurements for each sensor location, averaged over all lanes, computed from counts and occupancy data.

Fig. 3 represents the speed pseudo-measurements at each sensor location over time, constructed from the average vehicle counts and occupancies following (8). From this figure, given the relatively high density of sensors, one can identify for reference the times and locations of bottlenecks. However, these speed pseudo-measurements are not differentially-private and hence cannot be used directly in the EKF. Moreover they depend on the occupancy measurements, which can exhibit high sensitivity to a single vehicle trajectory.

Fig. 4 shows the density map reconstructed using a non-private EKF, i.e., using raw counts and occupancy measurements without sanitization. Hence this figure shows the baseline performance of the estimator, before the introduction of the privacy-preserving blocks of Fig. 2. The parameters of the fundamental diagram used for all algorithms are $v_f = 65$ mph, $w = 11.6$ mph, $\rho_J = 193$ vehicles/mile/lane.

Finally Fig. 5 shows an example of a $(\log(2)+M, 0.05)$-differential private map, where M is the number of sensors used over the spatial sub-interval of interest. The complete map is built by using 10 out of the 27 sensors, hence is $(10 + \log(2), 0.05)$ differentially private. However, if we were only interested in say the area between the postmiles 22.5 and 26.5, choosing 4 sensors in this area is enough to obtain a satisfying estimation quality,

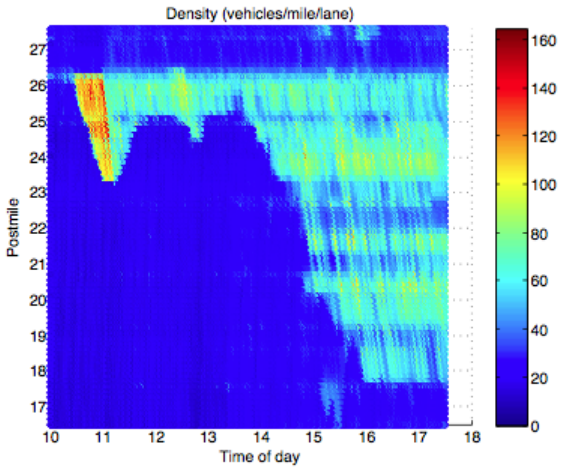

Figure 4: Real-time density map reconstruction with a non-private extended Kalman filter.

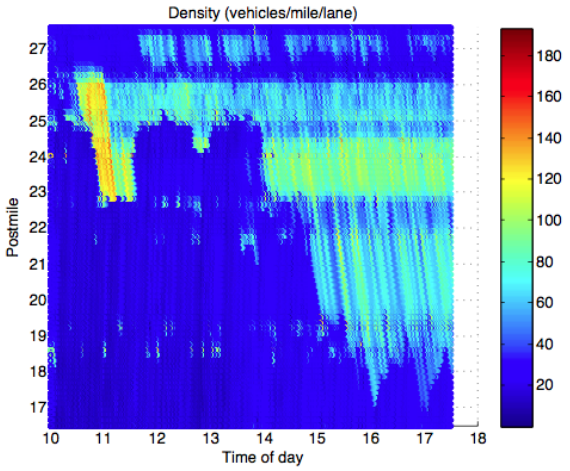

Figure 5: Real-time density map with $(\log(2) + M, 0.05)$-differential privacy guarantee, where M is the number of loop detectors used.

so publishing a map for this interval could be done with a $(4 + \log(2), 0.05)$-differential privacy guarantee.

A more thorough discussion of the effect of various parameter choices on the performance of the architecture is left for a full version of this paper.

7. CONCLUSION

This paper describes techniques that can guarantee the differential privacy of individual users whose data is used to provide online estimation of the traffic state on a road section. In contrast to previously proposed privacy-preserving schemes for location-based services, we specifically target the release of aggregated quantities, such as effective traffic speed and density, and we rely on a macroscopic hydrodynamic model of the dynamics of these variables to provide sufficiently accurate estimators. Future work will build on these ideas to develop privacy-preserving mechanisms for additional sensing modalities and complex road networks, and study how various choices of models and data assimilation mechanisms (e.g., Lagrangian vs. Eulerian) impact our ability to provide rigorous privacy guarantees.

Acknowledgments

This work was supported by NSERC under Grant RGPIN-435905-13.

8. REFERENCES

[1] E. S. Canepa and C. G. Claudel. A framework for privacy and security analysis of probe-based traffic information systems. In *Proceedings of the 2nd ACM international conference on High confidence networked systems (HiCoNS)*, pages 25–32, 2013.

[2] R. Chen, B. C. M. Fung, N. Mohammed, B. C. Desai, and K. Wang. Privacy-preserving trajectory data publishing by local suppression. In *Information Sciences (INS): Special Issue on Data Mining for Information Security*, volume 231, pages 83–97, May 2013.

[3] C. F. Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological*, 28(4):269–287, 1994.

[4] C. Dwork. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)*, volume 4052 of *Lecture Notes in Computer Science*. Springer-Verlag, 2006.

[5] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. *Advances in Cryptology-EUROCRYPT 2006*, pages 486–503, 2006.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Theory of Cryptography Conference*, pages 265–284, 2006.

[7] J. C. Herrera, D. B. Work, R. Herring, X. Ban, Q. Jacobson, and A. M. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568 – 583, 2010.

[8] B. Hoh, T. Iwuchukwu, Q. Jacobson, M. Gruteser, A. Bayen, J.-C. Herrera, R. Herring, D. Work,

M. Annavaram, and J. Ban. Enhancing privacy and accuracy in probe vehicle based traffic monitoring via virtual trip lines. *IEEE Transactions on Mobile Computing*, 11(5), May 2012.

[9] Z. Jia, C. Chen, B. Coifman, and P. Varaiya. The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors. In *Proceedings of the 4th IEEE Conference on Intelligent Transportation Systems*, 2001.

[10] J. Le Ny and G. J. Pappas. Differentially private filtering. *IEEE Transactions on Automatic Control*, 59(2):341–354, February 2014.

[11] M. Lighthill and G. Whitham. On kinematic waves II. A theory of traffic flow on long crowded roads. *Proceedings of the Royal Society*, 229A:317–345, May 1955.

[12] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *Proceedings of the IEEE Symposium on the Foundations of Computer Science*, 2007.

[13] L. Muñoz, X. Sun, R. Horowitz, and L. Alvarez. Traffic density estimation with the cell transmission model. In *Proceedings of the American Control Conference*, volume 5, pages 3750–3755. IEEE, 2003.

[14] N. Pham, R. K. Ganti, Y. S. Uddin, S. Nath, and T. Abdelzaher. Privacy-preserving reconstruction of multidimensional data maps in vehicular participatory sensing. In *Proceedings of the 7th European conference on Wireless Sensor Networks*, pages 114–130, 2010.

[15] P. I. Richards. Shockwaves on the highway. *Operations Research*, 4:42–51, 1956.

[16] L. Sankar, S. R. Rajagopalan, and H. V. Poor. A theory of privacy and utility in databases. Technical report, Princeton University, February 2011.

[17] R. Shokri, C. Troncoso, C. Diaz, J. Freudiger, and J.-P. Hubaux. Unraveling an old cloak: k-anonymity for location privacy. In *Proceedings of the CCS Workshop on Privacy in the Electronic Society (WPES)*, 2010.

[18] D. Simon. *Optimal State Estimation*. Wiley-Interscience, 2006.

[19] C. M. J. Tampere and L. H. Immers. An extended Kalman filter application for traffic state estimation using CTM with implicit mode switching and dynamic parameters. In *Proceedings of the IEEE Intelligent Transportation Systems Conference*, Seattle, WA, 2007.

[20] M. Treiber and A. Kesting. *Traffic Flow Dynamics*. Springer, 2013.

[21] Y. Wang and M. Papageorgiou. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transportation Research Part B: Methodological*, 39(2):141–167, February 2005.

[22] D. B. Work, O.-P. Tossavainen, S. Blandin, A. M. Bayen, T. Iwuchukwu, and K. Tracton. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In *Proceedings of the 47th IEEE Conference on Decision and Control*, pages 5062–5068. IEEE, 2008.

[23] Y. Yuan, J. W. C. van Lint, R. E. Wilson, F. van Wageningen Kessels, and S. P. Hoogendoorn. Real-time Lagrangian traffic state estimator for freeways. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):59–70, March 2012.

[24] H. Zhang and J. Bolot. Anonymization of location data does not work: a large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, 2011.