

Efficient Privacy-Preserved Processing of Multimodal Data for Vehicular Traffic Analysis

Meisam Mohammady
Iowa State University
meisam@iastate.edu

Reza Arablouei
Data61, CSIRO
reza.arablouei@csiro.au

Abstract—We estimate vehicular traffic states from multimodal data collected by single-loop detectors while preserving the privacy of the individual vehicles contributing to the data. To this end, we propose a novel hybrid differential privacy (DP) approach that utilizes minimal randomization to preserve privacy by taking advantage of the relevant traffic state dynamics and the concept of DP sensitivity. Through theoretical analysis and experiments with real-world data, we show that the proposed approach significantly outperforms the related baseline non-private and private approaches in terms of accuracy and privacy preservation.

I. INTRODUCTION

Differential privacy (DP) is commonly used in privacy-enhancing technologies, e.g., see [1]-[11] and the references therein. However, utilizing DP-based techniques can pose certain challenges in a variety of applications. Examples are applying DP to set-valued datasets, which requires using a context-free taxonomy tree [12], or to relational datasets, which is called non-interactive DP and usually done via sampling [13].

Datasets containing multimodal data appear in many real-world applications, e.g., regarding vehicular or network traffic or finance. Records in multimodal datasets are usually represented by value pairs such as $\langle x, y \rangle$ where x is the modal value, e.g., the number of cars in an area, and y is the mode, e.g., free or congested traffic. The straightforward application of DP to multimodal datasets requires randomizing both modal and mode values, which generally leads to significant sacrifice of accuracy.

In this paper, we present a novel hybrid DP approach that minimizes the required randomization through leveraging the underlying application-specific dynamics and the notion of sensitivity in DP, which is defined as the impact of changing the value of one data element over the outcome of a given query. We observe that multimodal dynamics often do not warrant randomizing the associated mode values as the DP sensitivity is not large enough to alter the mode. For instance, in vehicular traffic flows, adding or removing a single car can change the mode only at a transition point between the free to congested modes. Therefore, to apply DP in a traffic state

estimation application, we can define two zones, namely, the safe zone and the sensitive zone, where the former contains the traffic states where DP randomization is not required for the modes, and the latter contains the states that require randomizing the modes.

To illustrate the efficacy of the proposed hybrid DP approach, we consider a traffic monitoring application. The corresponding dataset is collected by 27 single-loop detectors installed at various locations of the US Interstate 80 highway. This dataset comprises 18 hours of $\langle \text{count}, \text{occupancy} \rangle$ pair values recorded over a day. The field “count” is the number of cars passing by a detector and “occupancy” is the time during which the sensor is activated. Each pair value is recorded over a 30s continuous time window and the occupancy values are given as fractions of the associated 30s periods. The problem of interest is specifically the estimation of traffic state from the observed data while preserving the privacy of individual vehicles. The dynamics of traffic state can be represented by a nonlinear stochastic state-space model where the state is the density of the cars on the road and the observation data is the flow of the traffic.

We verify the effectiveness of the proposed hybrid DP approach both theoretically and empirically by comparing its performance to that of three benchmark approaches. The considered benchmarks are (i) a baseline non-privacy-preserving (non-private) approach that utilizes an extended Kalman filter where its prediction block optimally computes trustworthy state values from count and occupancy values, (ii) a baseline DP approach that applies two randomization mechanisms, i.e., an additive Gaussian randomization of counts and an exponential randomization of mode values [14], and (iii) a variant of the proposed approach that applies exponential DP mechanism only at a defined sensitive zone.

We show that the privacy guarantee of the considered baseline DP approach depreciates as the number of sensors grows. Intuitively, more records aggregated by the sensors mandate exponentially stronger exponential DP mechanism for mode values to maintain the privacy guarantee due to the increase in the sensitivity value. Our proposed hybrid DP approach for multimodal data addresses this dependency through carefully examining the state dynamics and determining the sensitivity according to the traffic status.

II. TRAFFIC FLOW DYNAMICS

The unidirectional traffic along a single road section, with the position denoted by x and the varying number of lanes by

$\lambda(x)$, can be mapped based on the traffic flow dynamics [15] as $q = \rho v$ where ρ is the vehicle density (e.g., in vehicles per mile) over all lanes, q is the traffic flow over all lanes, and v is the traffic velocity. Here, we assume the simple case of a homogeneous road section [15] with the associated continuity equation expressed as

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho v)}{\partial x} = 0. \quad (1)$$

We consider a discrete version of (1) by dividing the road section into cells of length x_i and using a time step of t [15]. Hence, the density in cell i over all lanes follows the recursion

$$\rho^i(t + t) = \rho^i(t) + \frac{t}{x_i} [F_{\text{tot}}(\rho^{i-1}(t), \rho^i(t)) - F_{\text{tot}}(\rho^i(t), \rho^{i+1}(t))] \quad (2)$$

where $F_{\text{tot}}(\rho^{i-1}(t), \rho^i(t))$ is the total numerical flux that enters cell i (i.e., through the interface $i-1 \rightarrow i$) during period t , and $F_{\text{tot}}(\rho^i(t), \rho^{i+1}(t))$ is the total numerical flux out of cell i (i.e., through the interface $i \rightarrow i+1$). Note that the numerical flux $F_{\text{tot}}(\rho^i(t), \rho^{i+1}(t))$ is in general different from the total flow $q(x_{ij+1}, t)$, where x_{ij+1} denotes the location of the interface between cells i and $i+1$. More details are provided in the following.

To complete the model, we need to make a hypothesis on the relationship between two quantities, e.g., velocity and density, or flow and density. Thus, we first introduce lane-averaged (also called effective) quantities, i.e., 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 $q(x, t) = \rho(x, t)v(x, t)$ [15]. Denoting by $\rho_j(x, t)$, $q_j(x, t)$, and $v_j(x, t)$ the density, speed, and flow in lane j at position x , we have the relations

$$\begin{aligned} \rho(x, t) &= \frac{\sum_{j=1}^{(x)} \rho_j(x, t)}{\lambda(x)} \\ q(x, t) &= \frac{\sum_{j=1}^{(x)} q_j(x, t)}{\lambda(x)} \\ v(x, t) &= \frac{\sum_{j=1}^{(x)} v_j(x, t)}{\lambda(x)}. \end{aligned}$$

We adopt a first-order model or fundamental diagram, considering a static relationship $q(\rho)$. In first-order models, proposed in [16], [17] and known as Lighthill-Whitham-Richards (LWR) models, the effective density is a fundamental quantity and a sufficient description of the local traffic state, since the effective speed and the effective flow are assumed to be known static functions of density. The LWR models assume that the traffic flow is always in local equilibrium with respect to the density. This may lead to the formation of physically impossible phenomena such as shock waves. Regardless, the LWR models are widely used for modeling traffic flow dynamics [15]. Here, we use a triangular fundamental diagram as our LWR model. In the following, we describe this model, which is also called the cell-transmission model (CTM) [18].

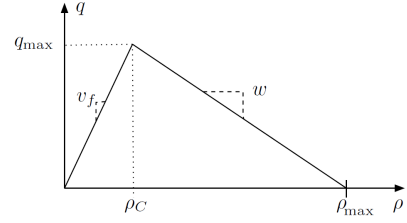


Fig. 1. Triangular fundamental diagram and the associated parameters.

A. Cell Transmission Model

The CTM is a simple LWR model that uses a triangular fundamental diagram formulated as

$$q(\rho) = \begin{cases} v_f \rho & \text{if } \rho \leq \rho_c \\ w(\rho_{\max} - \rho) & \text{if } \rho_c \leq \rho \leq \rho_{\max}. \end{cases} \quad (3)$$

Here, v_f is the velocity of free traffic (say 110 km/h for a highway), ρ_{\max} is the maximum density on this road segment (say 120 vehicles/lane/km for a highway), ρ_c is the critical density at which the maximum flow $q_{\max} = v_f \rho_c$ is attained, and w is the velocity of the waves of density variations in congested traffic that propagate backwards. Fig. 1 illustrates these definitions [15].

Dividing the road into I cells numbered as $1, \dots, I$, the discrete-time lane-averaged conservation law for vehicles corresponding to the solution $q(\rho)$ of (1) is

$$\rho_{k+1}^i = \rho_k^i + \frac{t}{x_i} \left(\frac{\lambda^{i-1}}{\lambda^i} F(\rho_k^{i-1}, \rho_k^i) - F(\rho_k^i, \rho_k^{i+1}) \right),$$

for $i = 1, \dots, I$, where ρ_k^i is the lane-averaged vehicle density in cell i at period k , i.e., during the time interval $[k t, (k+1) t]$, and $F(\rho_k^i, \rho_k^{i+1})$ is the lane-averaged numerical flux out of cell i , i.e., through the interface $i \rightarrow i+1$, during period k . We also define λ^i to be the number of lanes at the interface $i \rightarrow i+1$. Any location where the number of lanes changes is presumed to fall inside a cell. This leads to a system with non-linear (piecewise linear) dynamics. At the ends of the road for which we estimate the traffic, we add two ghost cells numbered 0 and $I+1$ to enforce the boundary conditions. In order to enforce the boundary conditions, we assume that there are loop detectors at the exit of cell 0 and at the entrance of cell $I+1$ [18]. Therefore, we obtain the following stochastic state-space model of the density dynamics on the road:

$$\rho_{k+1}^i = \rho_k^i + \frac{t}{x_i} \left(\frac{\lambda^{i-1}}{\lambda^i} F(\rho_k^{i-1}, \rho_k^i) - F(\rho_k^i, \rho_k^{i+1}) \right) + \gamma_k^i, \quad (4)$$

for $i = 1, \dots, I$. Here, γ_k^i is a Gaussian random variable whose variance can be tuned in the design of the state estimator, based on the relative confidence we place in the model or the observations. The dynamics of the ghost cells are also expressed as

$$\rho_{k+1}^0 = \rho_k^0 + \gamma_k^0, \quad \rho_{k+1}^{I+1} = \rho_k^{I+1} + \gamma_k^{I+1}. \quad (5)$$

Finally, for the triangular fundamental (3), the standard numerical method of Godunov corresponds to using the following numerical flux in (4) as

$$F(\rho_k^i, \rho_k^{i+1}) = \min(\rho_k^i v_f, \rho_c v_f, w(\rho_{\max} - \rho_k^{i+1})). \quad (6)$$

The stochastic state-space model (4) together with the measurements reported by single-loop detectors (see section III ahead) can be assimilated into an extended Kalman filter (EKF) to construct a traffic density map [22].

III. SINGLE-LOOP DETECTOR MEASUREMENT MODEL

The datasets provided by the flow sensors consist of sequences of counts $c_{j;k}^i$ and occupancies $o_{j;k}^i$ for $k \geq 0, 1 \leq i \leq S$, and $0 \leq j \leq \lambda^i$. Here, k indicates the related 30s period, S is the number of single-loop detectors reporting the records, and j is the lane number. The occupancy $0 \leq o_{j;k}^i \leq 1$ is a unitless number representing the fraction of the k th period during which any vehicle has passed before sensor i . The single-loop detectors cannot directly measure the traffic density or velocity at their locations. However, their measurements can be used to estimate these quantities. For a single-lane road equipped with single-loop detectors, the estimates are

$$v_j(t) \approx g \frac{c_j(t)}{o_j(t)T}, \quad q_j(t) \approx \frac{c_j(t)}{T}, \quad \rho_j(t) \approx \frac{o_j(t)}{g} \quad (7)$$

where T is the time period of the sensor (30 seconds here) and g is the g-factor, which denotes the average effective vehicle length at the sensor location and can vary over time. Similar to [19], to obtain a more robust approximation of density, we first compute the approximate flows based on the count data. That is, we express the flow ϕ_k^i around the sensor placed at the interface $i \rightarrow i+1$ for cells i and $i+1$ by the following non-linear measurement model

$$\phi_k^i = \frac{1}{\lambda^i T} \sum_{j=1}^i c_{j;k}^i = F(\rho_k^i, \rho_k^{i+1}) + \nu_k \quad (8)$$

where ν_k is a Gaussian random variable representing the measurement error or noise. We then define the density pseudo-measurement model as

$$z_k^i = z_k^{i+1} = \begin{cases} \frac{i}{v_f} & \text{if } m_k^i = F \\ \rho_{\max} - \frac{i}{w} & \text{if } m_k^i = C \end{cases} \quad (9)$$

where m_k^i denotes the traffic mode of the interface that is either free (F) or congested (C) corresponding to $\rho \leq \rho_c$ or $\rho > \rho_c$, respectively. This model can be obtained by inverting our triangular fundamental diagram presented in (3). The observation signal z_k^i is related to the density of the flow as

$$z_k^i = z_k^{i+1} = \rho_k^i + \eta_k^i = \rho_k^{i+1} + \eta_k^{i+1} \quad (10)$$

where η_k^i, η_k^{i+1} are assumed to be Gaussian random variables. This model requires determining the exact mode of the traffic flow. The strategy proposed in [19] is to use the reported occupancy measurements and estimate the traffic mode to be either fluid or congested based on whether $\frac{o_i}{g} \leq \rho_c$ or $\frac{o_i}{g} > \rho_c$, respectively.

The above measurements can result in frequent traffic mode estimation errors due to inaccurate approximation of g-factor. Given cars being at least 18 feet long and trucks being up to 60 feet long, the g-factor parameter is expected to range from 18 feet for inner car-only lanes to 60 feet in the early morning for outer lanes over fluid highways with heavy truck traffic. These mode measurements are difficult to handle from a DP point of view, because the occupancy time due to a single

vehicle, denoted by $\frac{l_v}{v_v}$ with l_v being the length and v_v the speed, can vary widely depending on its speed. As a result, the sensitivity of these occupancy measurements is high and the standard Gaussian perturbation mechanism exacerbates the reliability of the measurements, especially at low density [19]. We now present our mode measurement model, which takes both the occupancy and the count measurements into account to obtain a more reliable estimation of the traffic mode.

IV. NON-PRIVATE MODE AND DENSITY MEASUREMENTS

According to (9), two possible densities on the fundamental diagram correspond to each flow measurement $0 \leq \phi_k^i < q_{\max}$ (8). Based on (7), we can also form the lane-average contribution to the density via occupancy measurements as

$$y_k^i = \frac{1}{g\lambda^i} \sum_{j=1}^i o_{j;k}^i \quad (11)$$

The traffic mode pseudo-measurements can then be obtained as

$$M_k^i = M_k^{i+1} = \arg \min_{m_k^i} |z_k^i(m_k^i) - y_k^i|. \quad (12)$$

The model (12) estimates the mode as either free (F) or congested (C) based on which subfunction in the hybrid function $z_k^i(m_k^i)$ (9) is closer to the occupancy contribution to the density y_k^i . This model requires an accurate estimate of the g-factor parameter to guarantee that the minimum in (12) is correctly evaluated. This is challenging as g-factor can change over time and is generally not easy to estimate. To tackle this, we assume that g-factor is constant, namely 20 feet. We then bound the allowed deviation between the density pseudo-measurement z_k^i (3) and the occupancy contribution to density y_k^i (11). That is, for constant g , $\exists \zeta(g) > 0$ such that $|\log z_k^i - \log y_k^i| \leq \zeta(g) \forall i, k$. This limits the variations in g-factor as

$$\frac{1}{ge^{(g)}\lambda^i} \sum_{j=1}^i o_{j;k}^i \leq z_k^i \leq \frac{e^{(g)}}{g\lambda^i} \sum_{j=1}^i o_{j;k}^i \quad (13)$$

Given the assumed values of g and $\zeta(g)$, we define the sets

$$T_F = \left\{ (\phi_k^i, y_k^i) : \left| \log \frac{i}{v_f} - \log y_k^i \right| \leq \zeta(g) \forall i, k \right\} \quad (14)$$

$$T_C = \left\{ (\phi_k^i, y_k^i) : \left| \log \left(\rho_{\max} - \frac{i}{w} \right) - \log y_k^i \right| \leq \zeta(g) \forall i, k \right\} \quad (15)$$

corresponding to the flow ϕ_k^i satisfying our truncation in the free and congested modes, respectively. Consequently, we introduce the following lemma.¹

Lemma IV.1. *For any flow ϕ_k^i , defined in (8), we have*

$$\mathbf{1}_{T_F}((\phi_k^i, y_k^i)) \mathbf{1}_{T_C}((\phi_k^i, y_k^i)) = 1 \quad (16)$$

$$\text{iff } \phi_k^i \in \left[\frac{wv_f\rho_{\max}}{we^{2(g)} + v_f}, \frac{we^{2(g)}v_f\rho_{\max}}{w + e^{2(g)}v_f} \right]. \quad (17)$$

¹We provide the proofs of all theorems and lemmas in Appendix III of the extended version of this article archived at <https://arxiv.org/pdf/2302.09783.pdf>

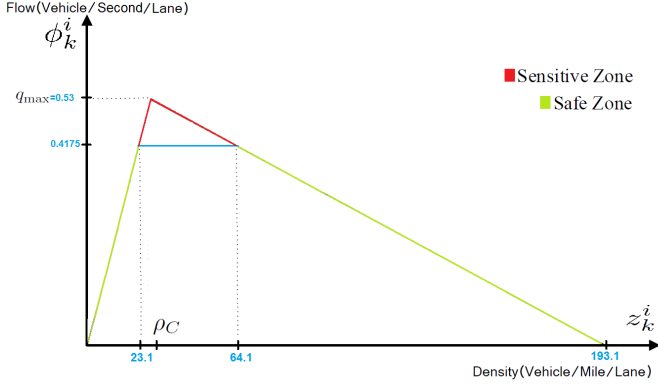


Fig. 2. Safe and Sensitive zones on a triangular fundamental diagram for $g = 20$ feet and $\zeta(g) = 0.51$.

Defining T_C and T_F , the complement sets of T_C and T_F , respectively, we obtain a traffic mode measurement model as

$$M_k^i = \begin{cases} F & \text{if } \mathbf{1}_{T_F} \setminus T_C((\phi_k^i, y_k^i)) = 1 : \text{Safe zone, F mode} \\ C & \text{if } \mathbf{1}_{T_C} \setminus T_F((\phi_k^i, y_k^i)) = 1 : \text{Safe zone, C mode} \\ M_k^i \text{ if } \left[\prod_{s=0}^{r-1} \mathbf{1}_{T_C \setminus T_F}((\phi_k^i \text{ }_s, y_k^i \text{ }_s)) \right] \times \\ \mathbf{1}_{T_C \setminus T_F}((\phi_k^i \text{ }_r, y_k^i \text{ }_r)) = 1, r > 0 : \text{Sensitive zone} \end{cases} \quad (18)$$

The mode measurement model (18) determines the mode of the traffic as either free (F) or congested (C), if the current flow satisfies (18) only in free mode or only in congested mode, respectively. The third case corresponds to the flow

$$\phi_k^i \in \left[\frac{wv_f \rho_{\max}}{we^2 \zeta(g) + v_f}, \frac{we^2 \zeta(g) v_f \rho_{\max}}{w + e^2 \zeta(g) v_f} \right]$$

where the truncation is respected in both traffic modes. For this case, we take the mode of the last flow $\phi_k^i \text{ }_r$, which is inside one of the two safe zones in (18). To illustrate this, we depict the region corresponding to the flows satisfying the truncation in one mode with green lines on a triangular fundamental diagram in Fig. 2. We call the region corresponding to these flows, the ‘‘Safe’’ zone. We also introduce a ‘‘Sensitive’’ zone (red line), which represents the flows that our model is unable to determine their mode directly. The parameters related to the fundamental diagram and the model that are used in Fig. 2 are $v_f = 65$ mph, $w = 11.6$ mph, $\rho_{\max} = 193$ vehicles/mile/lane, $g = 20$ feet, and $\zeta(g) = 0.51$. As per (13), the choice of $\zeta(g) = 0.51$ corresponds to g -factor variations between 12 and 33.3 feet.

Our mode measurement model (18) estimates the mode of the traffic with respect to the flows falling inside the Safe zone. Although this strategy can reduce the accuracy of the mode measurement, especially for flows in the Sensitive zone, in section V, we show that this model can efficiently be used to design a differentially-private traffic estimator.

To obtain a more physically meaningful mode, we filter the mode measurements through an additional hidden-Markov model (HMM), as explained below. For the state trajectory M_k^i defined in (18), we utilize the actual mode estimate used to

invert the fundamental diagram that is the new state trajectory $\{s_k^i\}_{k=0}$ with $s_k^i \in \{C, F\}$. We describe the dynamics of s_k^i via a Markov chain with a single parameter $\pi_1 = \mathbb{P}(s_{k+1}^i \neq s_k^i)$, which represents the probability of mode changing from free to congested at that location. This parameter can be estimated from historical data. We introduce another parameter $\pi_2 = \mathbb{P}(m_k^i = s_k^i)$ that reflects our confidence in the output of our model. We set the confidence probability parameter in HMM with respect to the flow data as its values for the flows in the Sensitive zone ought to be lower compared to those for the flows in the Safe zone. For non-private estimation, we can define the confidence probability as $\pi_2 = \mathbb{P}(m_k^i = s_k^i | q_k^i)$, which is useful for Sensitive zone flows. Even if $M_k^i \text{ }_r$ addresses a wrong mode, the confidence probability is set according to the occupancy contribution to density (11) hence HMM can correct the error. We summarize the procedure of providing density measurements from occupancies and counts data in Algorithm 1.

Algorithm 1 Non-private density measurement.

- 1: Calculate flow measurements $\phi_k^i = \frac{1}{T^i} \sum_{j=1}^i c_{j:k}^i$.
- 2: Based on historical data, choose a base g -factor, e.g, 20 feet, and an upper-bound error $\zeta(g)$.
- 3: Specify the corresponding Safe and Sensitive zones based on Theorem IV.1.
- 4: Calculate $m_k^i = \frac{F}{C}$ based on the mode measurement model (18).
- 5: Filter m_k^i through the HMM filter to obtain the actual mode s_k^i used to invert the fundamental diagram.
- 6: Calculate

$$z_k^i = z_{k+1}^i = \begin{cases} \frac{i}{v_f} & \text{if } s_k^i = F \\ \rho_{\max} - \frac{i}{w} & \text{if } s_k^i = C. \end{cases}$$

To illustrate our approach, we estimate the traffic state from induction loop data available as part of the Mobile Century dataset [20]. This data consists of counts and occupancy measurements from single-loop detectors for each northbound lane of US Interstate 880 highway between post-miles 16.5 and 27.7 (along an approximately 11-mile-long road section). We assimilate the density measurements based on Algorithm 1 in an EKF to construct the non-private density map shown in Fig. 3. The resulting map is similar to the map in Fig. 4 that is the non-private map produced by [19]. This similarity proves the reliability of our mode measurement model. The two maps have some discrepancies mainly in the areas where the traffic is about to switch between the modes. Fig. 3 is likely a more reliable picture of the traffic density as our mode measurement model considers the possible variations of g -factor over time. In the following, we show that this model can be used efficiently in a differentially-private scheme.

V. DIFFERENTIALLY-PRIVATE MODE AND DENSITY MEASUREMENTS

The measurements obtained from the single-loop detectors, i.e, counts $c_{j:k}^i$ and occupancies $o_{j:k}^i$, cannot be directly used in any traffic estimator architecture, because they may reveal private information about individuals who contribute to these measurements. In this section, we present

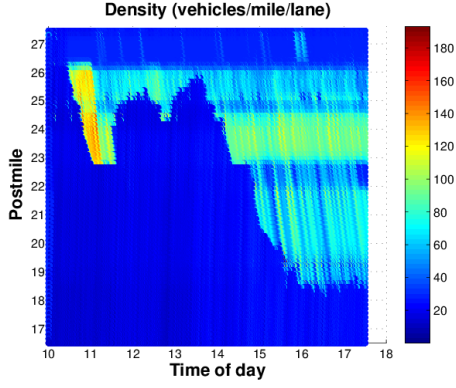


Fig. 3. Real-time density map reconstruction using a non-private EKF based on Algorithm 1.

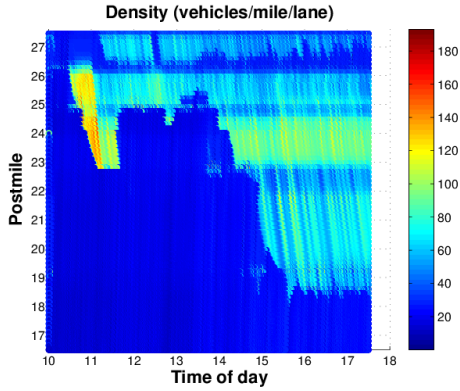


Fig. 4. Real-time density map reconstruction using a non-private EKF presented in [19].

deferentially private algorithms that output privacy-preserved flow and traffic mode measurements. These sanitized pseudo-measurements are then used to provide differentially-private density pseudo-measurements, which are sufficient to construct our differentially-private density map. We use a Gaussian mechanism, which was first presented in [19], to process the counts data and provide differentially-private traffic flow data.²

In order to preserve the privacy of traffic mode measurements, we introduce a new mechanism for sanitizing data sequences that is mainly based on Algorithm 1. In what follows, we first review the Gaussian mechanism providing the privacy-preserved flow measurements. Then, we present the mechanism that provides the mode pseudo-measurements.

A. Flow Measurements

We can consider the following adjacency relation for the counts datasets of N user trajectories $C := \{c_{j;k}^i : k \geq 0, 1 \leq$

$$i \leq S, 1 \leq j \leq \lambda^i\}$$

$$\forall c, e \in C : \text{Adj}(c, e) \text{ iff } \forall k \geq 0, \forall i \in [1, S], \forall j \in [1, \lambda^i], \\ i, j \in \mathbb{N}, \exists (j_1, k_1), (j_2, k_2) : |c_{j_1;k_1}^i - e_{j_1;k_1}^i| \leq 1, \\ |c_{j_2;k_2}^i - e_{j_2;k_2}^i| \leq 1, c_{j;k}^i = e_{j;k}^i \quad \forall (j, k) \neq (j_1, k_1), (j_2, k_2). \quad (19)$$

This adjacency relation indicates that changing the trajectory of a single car can affect the counts measurements reported by each sensor i in at most two different time steps. To make this clearer, suppose that Jane's car triggers a number of sensors every day when she goes to her job in the morning. For any DP mechanism to hide her absence or presence, we must keep in mind that changing her trajectory can change the reported counts of each sensor at two different times, one corresponding to a unit decrease in her usual trend of passing and the other corresponding to a unit increase in her new trend.

Let us consider two adjacent flow datasets as ϕ_k^i , expressed in (8), and $\tilde{\phi}_k^i$. Then, we have

$$\|\phi - \tilde{\phi}\|_2^2 = \sum_{i=1}^M \sum_{k=0}^7 |\phi_k^i - \tilde{\phi}_k^i|^2.$$

For a sensor at the interface $i \rightarrow i+1$, the corresponding term is

$$\sum_{k=0}^7 |\phi_k^i - \tilde{\phi}_k^i|^2 = \frac{1}{T^2(\lambda^i)^2} \sum_{k=0}^7 \left| \sum_{j=0}^i (c_{j;k}^i - e_{j;k}^i) \right|^2.$$

Based on the adjacency relation (19), the counts $c_{j;k}^i$ and $e_{j;k}^i$ must be almost all identical, except that some vehicles A and B can cross the line of the sensor at different periods and in different lanes. Thus, we have

$$\sum_{k=0}^7 |\phi_k^i - \tilde{\phi}_k^i|^2 \leq \frac{2}{T^2(\lambda^i)^2}$$

and hence

$$\|\phi - \tilde{\phi}\|_2^2 = \sum_{i=1}^M \sum_{k=0}^7 |\phi_k^i - \tilde{\phi}_k^i|^2 \leq \frac{2}{T^2} \sum_{i=1}^M \frac{1}{(\lambda^i)^2} : f^2. \quad (20)$$

The mechanism that publishes the perturbed flow pseudo-measurements $\tilde{\phi}_k^i = \phi_k^i + n_k^i$, where n_k^i are independent zero-mean white Gaussian noise with covariance κ^2 , f^2 and f as in (20), is (ϵ, δ) -differentially private [21].

B. Density and Mode Measurements

The flow pseudo-measurements obtained from the Gaussian mechanism, $\tilde{\phi}_k^i = \phi_k^i + n_k^i$, can be used to calculate the density pseudo-measurements, but this requires an additional mode estimate. However, estimating the traffic mode based on the count/occupancy datasets and without sanitization can compromise private information of individuals. In this section, we present our privacy-preserved mode measurement that is mainly based on the mode measurement model presented in Section IV. Like (19), the adjacency relation for the occupancy

²We provide some background information on the notion of DP in Appendix I of the extended version of this article archived at <https://arxiv.org/pdf/2302.09783.pdf>

data of N user trajectories $O := \{o_{j;k}^i : k \geq 0, 1 \leq i \leq S, 0 \leq j \leq \lambda^i\}$ is

$$\forall o, \theta \in O : \text{Adj}(o, \theta) \text{ iff } \forall k \geq 0, \forall i \in [1, S], \forall j \in [1, \lambda^i], \\ i, j \in \mathbb{N}, \exists (j_1, k_1), (j_2, k_2), \psi \in [0, 1] : |o_{j_1;k_1}^i - \theta_{j_1;k_1}^i| \leq \psi, \\ |o_{j_2;k_2}^i - \theta_{j_2;k_2}^i| \leq \psi, o_{j;k}^i = \theta_{j;k}^i \quad \forall (j, k) \neq (j_1, k_1), (j_2, k_2). \quad (21)$$

For the occupancy data, we bound the allowed deviation on the reported occupancy, when we add or remove one vehicle. This implies we offer no privacy protection for vehicles that change the measured cumulative occupancy or the average speed excessively (the occupancy contribution of one car is proportional to the inverse of its velocity). Therefore, the occupancy time due to a single vehicle is $o_{\text{car}}T = \frac{l_{\text{car}}}{v_{\text{car}}}$ where l_{car} is the average car length.

Considering the occupancy $0 \leq o_{j;k}^i \leq 1$, the adjacency relation (21) results in a high sensitivity and the corresponding standard Gaussian perturbation mechanism leads to unreliable occupancy pseudo-measurements, especially when the number of single-loop detectors in the road increases. Instead of using the occupancy measurements to estimate the density directly, we reconsider the mode pseudo-measurement model presented in Algorithm. 1 from a differential privacy perspective, that is, we examine how model (18) behaves when the trajectory of a single vehicle changes. Thus, let us rewrite (18) as

$$M_k^i = \begin{cases} F & \text{if } \mathbf{1}_{T_F} \tau_C \left(\binom{i}{k}, y_k^i \right) = 1 \\ C & \text{if } \mathbf{1}_{T_C} \tau_F \left(\binom{i}{k}, y_k^i \right) = 1 \\ M_{k_r}^i & \text{if } \left[\prod_{s=0}^{r-1} \mathbf{1}_{T_C \setminus T_F} \left(\binom{i}{k_s}, y_{k_s}^i \right) \right] \times \\ & \mathbf{1}_{T_C \setminus T_F} \left(\binom{i}{k_r}, y_{k_r}^i \right) = 1, r > 0 \end{cases} \quad (22)$$

by replacing the flow measurements ϕ_k^i with the flow pseudo-measurements \hat{y}_k^i . By changing the trajectory of a single vehicle, we have

$$M_k^i = \begin{cases} F & \text{if } \mathbf{1}_{T_F} \tau_C \left(\binom{i}{k}, \hat{y}_k^i \right) = 1 \\ C & \text{if } \mathbf{1}_{T_C} \tau_F \left(\binom{i}{k}, \hat{y}_k^i \right) = 1 \\ M_{k_r}^i & \text{if } \left[\prod_{s=0}^{r-1} \mathbf{1}_{T_C \setminus T_F} \left(\binom{i}{k_s}, \hat{y}_{k_s}^i \right) \right] \times \\ & \mathbf{1}_{T_C \setminus T_F} \left(\binom{i}{k_r}, \hat{y}_{k_r}^i \right) = 1, r > 0. \end{cases} \quad (23)$$

Defining $\hat{y}_k^i - y_k^i = \tilde{y}_k^i$, $\tilde{y}_k^i - \hat{y}_k^i = \bar{y}_k^i$, and according to the adjacency relations defined in (19) and (21), we have

$$\forall i \in [1, S], \exists k_1, k_2 : \left| \tilde{y}_{k_1}^i \right| \leq \frac{\psi}{g\lambda^i}, \left| \bar{y}_{k_2}^i \right| \leq \frac{\psi}{g\lambda^i}, y_k^i = 0, \\ \left| \hat{y}_{k_1}^i \right| \leq \frac{1}{T\lambda^i}, \left| \hat{y}_{k_2}^i \right| \leq \frac{1}{T\lambda^i}, \hat{y}_k^i = 0 \quad \forall i \neq i_0. \quad (24)$$

Consequently, we characterize all possible cases of mode switching due to changing the trajectory of a single vehicle in the following lemma.

Lemma V.1. *For the sets T_F and T_C defined in (14) and (15), and all flows \hat{y}_k^i , we have*

$$\mathbf{1}_{T_F} \left(\binom{i}{k}, \hat{y}_k^i \right) \mathbf{1}_{T_C} \left(\binom{i}{k}, \hat{y}_k^i \right) = 0 \\ \& \mathbf{1}_{T_C} \left(\binom{i}{k}, \hat{y}_k^i \right) \mathbf{1}_{T_F} \left(\binom{i}{k}, \hat{y}_k^i \right) = 0, \text{ if } \hat{y}_k^i \notin [\alpha, q_{\max}] \quad (25)$$

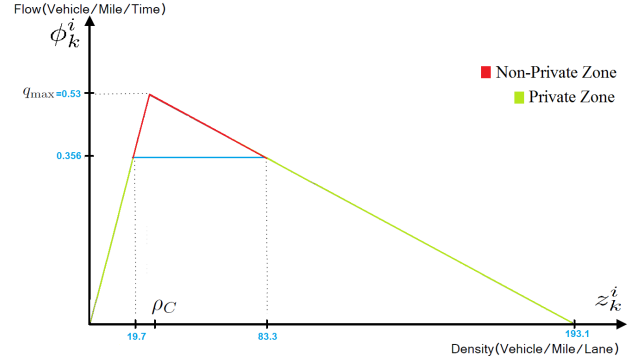


Fig. 5. The traffic mode in the Private zone is robust to any change in the trajectory of a single vehicle.

where

$$\alpha = \min \left\{ \frac{e^{-g} \left(\rho_{\max} - \frac{1}{T^i w} \right) - g^{-i}}{\frac{e^{\zeta(g)}}{v_f} + \frac{1}{e^{\zeta(g)} w}}, \frac{e^{-g} \rho_{\max} - \frac{e^{\zeta(g)}}{T^i v_f} - g^{-i}}{\frac{e^{\zeta(g)}}{v_f} + \frac{1}{e^{\zeta(g)} w}} \right\}.$$

This lemma shows that, for the flow pseudo-measurement $\hat{y}_k^i \leq \alpha$, changing the trajectory of a single vehicle does not affect the outcome of mode measurement model (22). Accordingly, we now develop our privacy-preserved mode measurement model. Based on Lemma V.1, we first define the sets

$$PT_F = \left\{ \left(\binom{i}{k}, y_k^i \right) : \left| \log \left[\frac{\hat{y}_k^i}{v_f} \right] - \log [y_k^i] \right| \leq \zeta(g), \right. \\ \left. \hat{y}_k^i \in [0, \alpha] \quad \forall i, k \right\}$$

$$PT_C = \left\{ \left(\binom{i}{k}, y_k^i \right) : \left| \log \left[\rho_{\max} - \frac{\hat{y}_k^i}{w} \right] - \log [y_k^i] \right| \leq \zeta(g), \right. \\ \left. \hat{y}_k^i \in [0, \alpha] \quad \forall i, k \right\}$$

corresponding to the pseudo-flow \hat{y}_k^i satisfying our private truncation in free or congested modes. Defining PT_C and PT_F as the complement sets of PT_C and PT_F , respectively, we obtain the following privacy-preserved mode measurement model:

$$M_k^i = \begin{cases} F & \text{if } \mathbf{1}_{PT_F} \tau_C \left(\binom{i}{k}, \hat{y}_k^i \right) = 1 \\ C & \text{if } \mathbf{1}_{PT_C} \tau_F \left(\binom{i}{k}, \hat{y}_k^i \right) = 1 \\ M_{k_r}^i & \text{if } \left[\prod_{s=0}^{r-1} \mathbf{1}_{PT_C \setminus PT_F} \left(\binom{i}{k_s}, \hat{y}_{k_s}^i \right) \right] \times \\ & \mathbf{1}_{PT_C \setminus PT_F} \left(\binom{i}{k_r}, \hat{y}_{k_r}^i \right) = 1, r > 0 \end{cases} \quad (26)$$

As shown in Fig. 5, we can divide the triangular fundamental diagram into two zones, called the ‘‘Private’’ zone and the ‘‘Non-Private’’ zone. For illustrative purposes, we depict the lines corresponding to these zones with different colors in Fig. 5. The Private zone represents the flow intervals for which the traffic mode can be estimated uniquely, and also changing the trajectory of a single vehicle does not affect the mode estimation. The diagram is depicted for a four-lane road and the parameters related to the fundamental diagram and the model are $v_f = 65$ mph, $w = 11.6$ mph, $\rho_{\max} = 193$

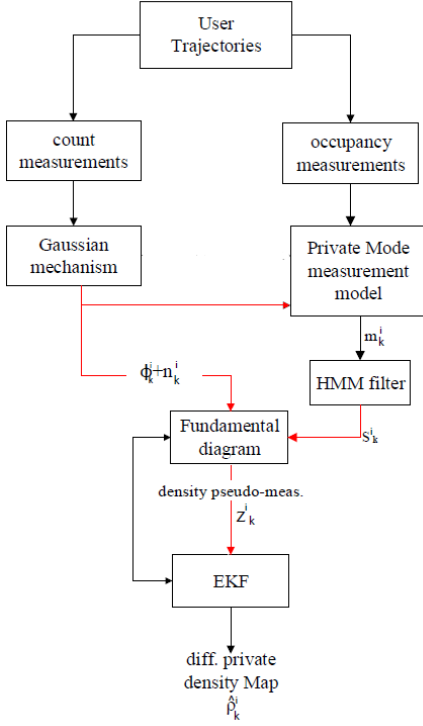


Fig. 6. Architecture of our differentially-private traffic estimator. The red arrows represent differentially-private signals, i.e., perturbed flow pseudo-measurements from vehicle counts. The private mode estimation is done using both counts and occupancy measurements.

vehicles/mile/lane, $g = 20$ feet, and $\zeta(g) = 0.51$. We also set $\psi = 0.25$ that is sufficiently large to protect the privacy of individuals, i.e., all the vehicles that cross the sensor line faster than 3 km/h, assuming the vehicles are at least 7 meters long. Our model for flows in the Non-Private zone estimates the traffic mode at each sensor location based on the previous estimated mode for which the flow is in the Private zone. Adopting this strategy minimizes the possibility of privacy leakage while it still provides a meaningful observation signal to specify the mode.

VI. TRAFFIC STATE ESTIMATION

In this section, we present the overall architecture of our differentially-private traffic state estimator, and illustrate its performance on the Mobile Century dataset [20]. Fig. 6 illustrates the overall architecture of our privacy-preserving traffic estimator. The EKF assimilates the dynamic traffic model (4) and the density pseudo-measurements z_k^i obtained from the occupancy and count measurements.

The differential privacy guarantee provided by this architecture is the sum of guarantees provided by the Gaussian mechanism and our privacy-preserving mode measurement model. We specify the privacy guarantee provided by our mode measurement model in the following lemma and subsequent theorem.

Lemma VI.1. Consider $D = \{c_{1;k}^i, \dots, c_{i;k}^i, o_{1;k}^i, \dots, o_{i;k}^i \mid \forall i, k\}$ collecting the count and occupancy data with adjacency relations defined by (19) and 21. Let d, d^0 be two adjacent

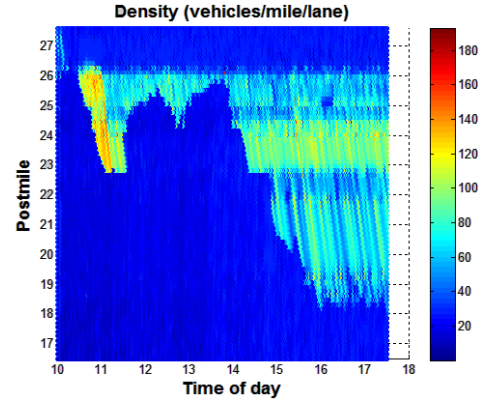


Fig. 7. Real-time density map reconstruction with $(\log(2);0.05)$ -DP guarantee presented based on our approach.

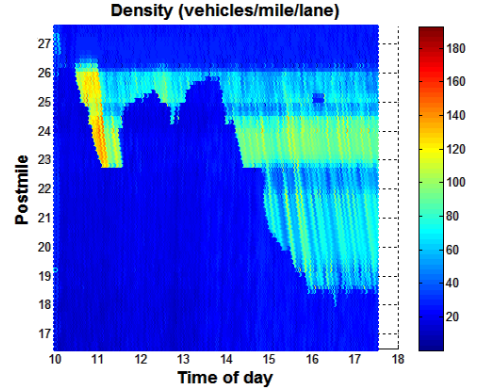


Fig. 8. Real-time density map reconstruction with $(\log(4);0.1)$ -DP guarantee presented based on our approach.

elements in D with K rows and define $E = \{E_i : E_i = e_1 \times e_2 \times \dots \times e_K, e_j \in [0, \alpha] \text{ or } [\alpha, q_{\max}] \forall i = 1, \dots, 2^K\}$ with α defined in Lemma V.1 specifying the Private zone. Then, for our privacy-preserving mode measurement mechanism M and the flow pseudo-measurement z_k^i , we have

$$M(d) = M(d^0) \text{ if } (d), (d^0) \in E_i, \forall d, d^0 \in D, \forall i = 1, \dots, 2^K. \quad (27)$$

Theorem VI.2. The privacy-preserving mode estimation mechanism defined in (26) is (ϵ, δ) -differentially private.

Finally, in light of the compositional property of DP, the DP guarantee for the overall architecture is $(2\epsilon, 2\delta)$, the aggregation of (ϵ, δ) -DP for the mode measurement and (ϵ, δ) -DP for the flow pseudo-measurement.

VII. RESULTS AND DISCUSSION

Figs. 7 and 8 show examples of $(\log(2), 0.05)$ and $(\log(4), 0.1)$ -differentially-private maps, respectively, based on our designed privacy-preserving traffic estimator. The complete map is built using 10 out of the 27 sensors placed at different locations on four lanes of the US I-880 highway.

Our proposed approach improves over the state-of-the-art in terms of privacy guarantee and has immediate applications in providing privacy preservation for traffic monitoring in long highways. It also improves the reliability of the reconstructed

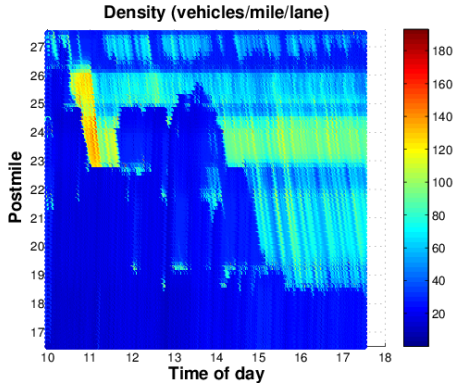


Fig. 9. Real-time density map reconstruction with $(10 + \log(2); 0.05)$ -DP guarantee presented in [19].

density maps by significantly reducing the instances of incorrect switching between the modes. Comparison of the three maps in Figs. 7, 8, and 3 shows that we can obtain strong $(\log(4), 0.1)$ or even stronger $(\log(2), 0.05)$ privacy guarantees using the proposed approach with negligible degradation in estimation performance. However, our approach may estimate the mode of the flows in the Non-Private zone with error, mainly due to the delay introduced by our mode measurement model. For example, assume that the traffic flow increases and the road becomes congested. The mode of the traffic based on our privacy-preserving mode measurement model will be free until the flow reenters the Private zone. In this case, according to Fig. 5, we may encounter up to 60 (vehicles/mile/lane) errors in our density maps. The upper-bound of these errors can be tightened by decreasing the value of the parameter ψ . For example, $\psi = 0.1$ can decrease the upper-bound of these errors to 40 units, but it also weakens the privacy guarantee. One effective way to improve the mode measurement when the flow is in the Sensitive zone is to use this model:

$$\forall i_k \in [\alpha, q_{\max}] : \begin{cases} F & \text{if } i_k - i_{k-1} > 0 \\ C & \text{if } i_k - i_{k-1} < 0. \end{cases} \quad (28)$$

This model takes advantage of the fact that the traffic flow dynamic is either decreasing or increasing given the congested or free modes, respectively. However, at this point, it unfortunately appears unwieldy for use with a more advanced mechanism.

VIII. CONCLUSION

We presented new methods for monitoring traffic while protecting the privacy of individual users whose data is used to estimate the traffic state at a particular location of a road. To this end, we used a macroscopic hydrodynamic model to analyze the dynamics of the variables involved. Our approach was different from earlier privacy-preserving methods used for location-based services as we focused on releasing aggregated data, such as traffic speed and density, while ensuring that the privacy of individuals is not compromised.

ACKNOWLEDGMENT

We thank Dr Jerome Le Ny for providing invaluable guidance and support throughout this work.

REFERENCES

- [1] G. Acs, C. Castelluccia, and R. Chen. 2012. Differentially private histogram publishing through lossy compression. In *12th IEEE Int. Conf. Data Mining*. Brussels, Belgium, 1–10.
- [2] H. Brenner and K. Nissim. 2010. Impossibility of differentially private universally optimal mechanisms. In *IEEE Annual Symp. Foundations of Computer Science*. Las Vegas, Nevada, USA, 71–80.
- [3] J. W. S. Brown, O. Ohrimenko, and R. Tamassia. 2013. Haze: Privacy-preserving real-time traffic statistics. In *Proc. ACM Int. Conf. Advances in Geographic Information Systems*. ACM, New York, USA, 540–543.
- [4] T. Chanyaswad, A. Dytso, H. Vincent Poor, and P. Mittal. 2018. MVG mechanism: Differential privacy under matrix-valued query. In *Proc. ACM Conf. Computer and Communications Security*. Toronto, ON, Canada, 230–246.
- [5] C. Dwork and A. Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* 9, 3–4, 211–407.
- [6] M. Hardt, K. Ligett, and F. McSherry. 2012. A simple and practical algorithm for differentially private data release. In *Proc. Conf. on Neural Information Processing Systems*. Lake Tahoe, NV, USA, 2348–2356.
- [7] A. Inan, M. Kantarcioglu, G. Ghinita, and E. Bertino. 2010. Private record matching using differential privacy. In *Proc. Int. Conf. Extending Database Technology*. 123–134.
- [8] C. Li, G. Miklau, M. Hay, A. McGregor, and V. Rastogi. 2015. The matrix mechanism: optimizing linear counting queries under differential privacy. *VLDB J.* 24, 6 (2015), 757–781.
- [9] I. Mironov. 2017. Rényi differential privacy. In *IEEE Computer Security Foundations Symp.*. IEEE, 263–275.
- [10] A. Nikolov, K. Talwar, and L. Zhang. 2013. The geometry of differential privacy: the sparse and approximate cases. In *Symp. Theory of Computing Conference*. Palo Alto, CA, USA, 351–360.
- [11] V. Rastogi and S. Nath. 2010. Differentially private aggregation of distributed time-series with transformation and encryption. In *Proc. ACM Int. Conf. Management of Data*. Indianapolis, IN, USA, 735–746.
- [12] R. Chen, N. Mohammed, B. Fung, B. Desai, and L. Xiong. 2011. Publishing set-valued data via differential privacy. In *Proc. VLDB Endow.* 4, 11, Aug. 2011, 1087–1098.
- [13] D. Leoni. 2012. Non-interactive differential privacy: a survey. In *Proc. Int. Workshop on Open Data*. New York, NY, USA, 40–52.
- [14] J. Le Ny and M. Mohammady. 2018. Differentially private MIMO filtering for event streams. *IEEE Trans. Automat. Control* 63, 1, 145–157.
- [15] M. Treiber, A. Kesting. 2013. *Traffic Flow Dynamics: Data, Models and Simulation*. Springer, 2013.
- [16] M. J. Lighthill and G. B. Whitham. 1955. On kinematic waves. II. A theory of traffic flow on long crowded roads. In *Proc. Royal Society of London A: Mathematical, Physical and Engineering Sciences*. vol. 229, no. 1178, pp. 317–345, The Royal Society, May 1955.
- [17] P. I. Richards. 1956. Shock waves on the highway. *Operations Research* 4(1), 42–51, 1956.
- [18] 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), pp. 269–287, Aug. 1994.
- [19] J. Le Ny, A. Touati, and G. Pappas. 2014. Real-time privacy-preserving model-based estimation of traffic flows. In *Proc. ACM/IEEE Int. Conf. Cyber-Physical Systems*. Apr. 2014, 92–102.
- [20] <https://traffic.berkeley.edu/project/mobilecentury>
- [21] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *Int. Conf. Theory and Applications of Cryptographic Techniques*. Springer, Berlin, Heidelberg, 486–503.
- [22] D. Simon. 2006. *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley-Interscience, 2006.