On the Privacy of Crowd-sourced Data Collection for Distance-to-Empty Prediction and Eco-routing

Chien-Ming Tseng
Masdar Institute of Science and Technology
ctseng@masdar.ac.ae

Chi-Kin Chau
Masdar Institute of Science and Technology
ckchau@masdar.ac.ae

ABSTRACT

The paradigm of crowd-sourced data collection (also known as participatory sensing) has been bolstered by the extensive availability of on-board sensors and electronic devices in nowadays vehicles, which can be applied in a wide range of transportation applications. Distance-to-empty (DTE) is the distance an electric vehicle (EV) or internal-combustion engine (ICE) vehicle can reach before its battery/fuel is exhausted, which is determined by a variety of uncertain factors, such as driving behavior, terrain, types of road, traffic, and vehicle specification. Eco-routing aims to optimize the route selection with lower energy consumption. The accuracy of DTE prediction and eco-routing can be enhanced substantially by the crowd-sourced data collected from diverse drivers and vehicles. However, a critical concomitant issue for crowd-sourced data collection is privacy, because the personal travel history may be misused without consents from the contributing users. To encourage large-scale adoption and contributions of crowd-sourced data collection from end users, this paper addresses the issue of privacy and proposes possible solutions to tackle the challenges. In particular, we discuss a solution of matrix factorization from collaborative filtering to enhance the privacy of crowd-sourced data collection in the context of transportation applications, such as DTE prediction and eco-routing.

CCS Concepts

•**Applied computing** → *Transportation;* •**Security and privacy** → *Privacy-preserving protocols;*

Keywords

Privacy, Crowd-sourced Data Collection, Energy-efficient Transportation

1. INTRODUCTION

In crowd-sourced data collection (also known as participatory sensing), a group of users contribute their personal data (possibly, voluntarily) to a third-party data repository, in exchange for the useful knowledge extracted from the collective data. The knowledge will be incorporated in personalized applications of individual users. Crowd-sourced data collection has been applied in diverse applications of pervasive and mobile computing systems [3]. Recently, vehicles are becoming a vital platform for crowd-sourced data collection. First, there are extensive deployments of on-board sensors and in-vehicle information systems, equipped with network connectivity and computing power, acting as effective data collection systems. Second, the wide availability of electronic devices and smartphones carried by passengers can extend the sensing abilities of vehicles. Third, there are abundant off-the-shelf and after-market automotive accessories for gathering driving data and vehicle information. Notably, crowd-sourced data collection for vehicles has been applied in certain existing transportation applications (e.g., traffic status updates in Google Map and Waze).

Several driving energy efficiency applications can be enhanced substantially by crowd-sourced data. One of the critical applications is the prediction of *distance-to-empty* (DTE) - the distance an electric vehicle (EV) or internal-combustion-engine (ICE) vehicle can reach before its battery/fuel is exhausted. DTE is determined by a variety of factors, such as driving behavior, terrain, types of road, traffic, and vehicle specifications. The conventional approach of DTE prediction employed by car manufacturers is based on the projection of past average vehicle energy efficiency of individual drivers. Such an approach is often perceived to be inaccurate. However, if there is sufficient knowledge about the vehicle, driving behavior and the route to travel, future energy efficiency can be estimated with higher accuracy. The availability of crowd-sourced data is able to improve the accuracy of DTE prediction by exploiting the historical data from other drivers. Conceptually, one can identify the factors pertaining to various types of drivers, vehicles or environments. Then, one can interpolate the data from similar drivers, vehicles or environments to assist the prediction. A framework of crowd-sourced data collection with appropriate factor extraction and incorporation mechanisms for personalized applications can be applied to many other aspects, such as eco-routing, driving coaching and refueling planning (as depicted in Fig. 1).

There are two common approaches of incorporation of crowd-sourced data collection in personalized applications for vehicles:

1. *Comparison with the Average*: One can obtain the global characteristics by the average data values (e.g.,

EV-Sys, June 21-24 2016, Waterloo, ON, Canada

ISBN 978-1-4503-4420-3/16/06…$15.00

DOI: http://dx.doi.org/10.1145/2939953.2939956

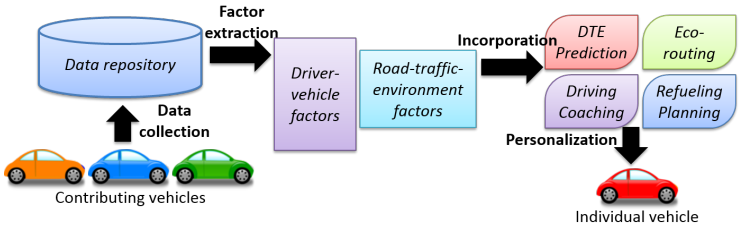

Figure 1: A framework of crowd-sourced data collection for vehicles.

average speed, mean stopping duration) from a large crowd-sourced dataset for a specific environment. Then we compensate the deviation of individual drivers from the average data values in personalized applications.

2. *Collaborative Filtering*: A domain-free data mining technique analyzes the local relationships and similarities within a dataset to identify a smaller set of factors or features that can characterize the observed data. The major approaches of collaborative filtering are neighborhood methods and latent factor models [11]. In particular, matrix factorization is a primary solution of the latent factor models.

While crowd-sourced data collection can provide various driving and vehicle data, a critical concomitant issue is privacy. The personal travel history may be misused without consents from the contributing users, which discourages the adoption and contributions from end users. This paper address the issues of privacy and proposes possible solutions to tackle the challenges. It is vital to address the issue of privacy in the common approaches of crowd-sourced data collection for vehicles. In this paper, we particularly discuss a solution of matrix factorization from collaborative filtering to enhance the privacy of crowd-sourced data collection in the context of transportation applications, such as DTE prediction and eco-routing.

2. RELATED WORK

2.1 Data Collection from Vehicles

The accuracy of driving energy consumption prediction can be enhanced by extensive data collection. Two crucial factors of energy consumption prediction are the future speed profiles and future environmental factors (e.g., temperature, wind speed or route grade), which may be highly dynamic and difficult to predict. Also, one can deploy sensor networks, by which stationary measurements at specific locations, such as traffic, average speed, speed limit and route grade can be measured. There are a number of papers focusing on utilizing such information [2, 8, 10]. A study that integrates the real-time traffic sensor data to predict the energy consumption and emission of ICE vehicles is presented in [17]. One can also obtain the estimated information from social networks and participatory sensing. Participatory sensing can provide mobile measurements and good geographic penetrations [16]. For example, a study shows that the estimation of stochastic effects which impact the travel velocity and acceleration profiles can be crowd-sourced to identify traffic congestion [5]. Our previous work has employed participatory sensing and provides personalized DTE prediction system based on participatory sensing data [13–15].

2.2 Privacy Aware Data Collection

The notion of ensuring privacy in data collection has been extensively studied in the literature of privacy and security. In particular, there are two common approaches: (1) obfuscation by incorporating random perturbation, and (2) cryptographic techniques such as multi-party computation. Privacy-preserving data mining is a popular topic of incorporating random perturbation to prevent the inference of individual users' information, while still allowing meaningful data mining queries. There are several recent studies [1, 9] that present mechanisms for differential privacy in matrix factorization using Laplace mechanisms by incorporating artificial random noise. On the other hand, cryptographic approach, such as garbled circuit, has been employed in matrix factorization [12], which requires higher computational overhead. These studies consider the primary applications for recommender systems, while this paper focus on the context of transportation applications, such as DTE prediction and eco-routing.

3. SETTINGS OF PRIVACY

The notion of privacy is related to the ability of seclusion of ones' private information through carefully designed interactions with the external world. In the context of transportation applications, individual drivers may be cautious about revealing their personal information (e.g., travel history, driving habits, vehicle types, etc.) during the course of crowd-sourced data collection process.

Privacy is not a binary option. There are different granularities and threat models under consideration. We first discuss different levels of privacy settings in transportation applications.

- **Complete Privacy**: No outsider can learn any individual's data from the public data repository. Outsiders cannot identify an individual who has contributed personal data to the data repository, nor their types.
- **Type-revealing Privacy**: Outsiders may infer general type-based information of the individuals who have contributed personal data to the public data repository (e.g., driving habits or vehicle types). But outsiders cannot infer an individual's fine-grained data (e.g., the detailed travel history and locations).
- **Identity-revealing Privacy**: Outsiders may infer an individual who has contributed personal data to the public data repository, by associating with their contributed data. But outsiders cannot infer an individual's fine-grained data (e.g., the detailed travel history and locations).

If providing complete privacy is too difficult or impractical, it would suffice to offer type-revealing or identity-revealing privacy in practice, in particular, to balance the trade-off between the usefulness of collected data and the privacy of contributed users. Note that type-revealing or identity-revealing privacy is sometimes acceptable in transportation applications, for example, certain vehicle-owner information is already available in public databases.

Next, we discuss the threat models related to the trustworthiness of data collector who processes individuals' contributed data to release to the data repository.

- **Trusted Data Collector**: End users can reveal their identities and personal travel histories to the data collector. But the data collector needs to ensure the privacy in the public data repository.
- **Untrusted Data Collector**: End users do not reveal their identities nor personal travel histories to the data collector. The communication between users and the data collector may be conducted in an anonymous medium. Any information revealed by the users will be released to the public data repository.

4. MODEL

In this section, we first present the driving energy consumption models used for crowd-sourced data collection. We will discuss the mechanisms to ensure privacy in the crowd-sourced data collection process in the subsequent section.

4.1 Areas of Factors

While there are many factors to determine driving energy consumption, they can be classified by three broad areas:

- **Driver**: The driver who controls the vehicle has a direct impact on the vehicle movement. Different drivers exhibit different preferences for stop/start and acceleration, aggression in various scenarios, propensity for hypermiling, etc. Psychological and behavioral traits of drivers also affect driving energy efficiency.
- **Vehicle**: Different types of vehicles consume energy differently. ICE vehicles are characterized by the engine types and gear shifts, whereas hybrid and EVs are affected by battery performance and regenerative braking. The sizes and weights of vehicles often determine the efficiency of kinetic energy conversion, so SUVs and trucks are usually less energy-efficient than sedan and compact vehicles.
- **Environment**: The environmental factors include both traffic and roads. Traffic for a road segment depends on a plethora of factors, including time-of-day, day-of-year, special events, which may follow a certain pattern. The types of roads also affect drivers' behavior differently, which can be divided into three main categories: small public or private roads with urban traffic, lower capacity "urban" highways, and higher capacity freeways.

The historical data of vehicle speed profiles can be identified by a combination of (driver, vehicle, environment), referred as a *data point*. Through crowd-sourced data collection, a dataset of measured energy consumption for a relatively small number of data points are collected. We will interpolate the missing data points from the collected data points.

4.2 Energy Consumption Model

This section describes a linear blackbox model of driving energy consumption that has been used extensively in the literature [4,8,14]. We denote a driver by D, a vehicle model by V, and a particular environment (e.g., a segment of route and time-of-day) by R. Each energy consumption is represented by a numerical value $E_{\mathsf{D},\mathsf{V},\mathsf{R}}$, indexed by the tuple $(\mathsf{D},\mathsf{V},\mathsf{R})$. All the entries of energy consumption values form a 3-dimensional tensor, denoted by $[E_{\mathsf{D},\mathsf{V},\mathsf{R}}]$.

While there are sophisticated approaches of estimating the moving vehicle energy consumption by white-box microscopic behavior models, these models are rather difficult to implement. Many parameters are required, for example, engine efficiency, transmission efficiency, regenerative braking efficiency, etc. However, in practice, these parameters are hard to obtain. Therefore, this paper utilizes a blackbox approach without the detailed knowledge of vehicle mechanics.

The total energy consumption E of driver D with vehicle model V in a particular environment R is given by:

$$E_{\mathsf{D},\mathsf{V},\mathsf{R}} = E^{\text{mv}}_{\mathsf{D},\mathsf{V},\mathsf{R}} + E^{\text{id}}_{\mathsf{D},\mathsf{V},\mathsf{R}} \tag{1}$$

where $E^{\text{mv}}_{\mathsf{D},\mathsf{V},\mathsf{R}}$ is the moving vehicle energy consumption and $E^{\text{id}}_{\mathsf{D},\mathsf{V},\mathsf{R}}$ is the idle vehicle energy consumption. We next present simple blackbox models to characterize $E^{\text{mv}}_{\mathsf{D},\mathsf{V},\mathsf{R}}$ and $E^{\text{id}}_{\mathsf{D},\mathsf{V},\mathsf{R}}$.

4.2.1 Moving Vehicle Energy Consumption

With respect to a particular combination of $(\mathsf{D},\mathsf{V},\mathsf{R})$, the moving vehicle energy consumption E^{mv} has unit in liter or kWh. Next, we drop the subscript $_{\mathsf{D},\mathsf{V},\mathsf{R}}$ for brevity.

In this paper, we estimate E^{mv} (denoted by $\hat{E}^{\text{mv}}$) by a linear equation of several measurable variables from vehicles[1]:

$$\hat{E}^{\text{mv}} = \begin{bmatrix}\alpha_{v,1}\\ \alpha_{v,2}\\ \vdots\\ \alpha_{v,r}\end{bmatrix}\begin{bmatrix}v\\ v^2\\ \vdots\\ v^r\end{bmatrix}^T + \begin{bmatrix}\vec{\alpha}_{d,1}\\ \vec{\alpha}_{d,2}\\ \vdots\\ \vec{\alpha}_{d,k}\end{bmatrix}\begin{bmatrix}\vec{d}\\ \vec{d}^2\\ \vdots\\ \vec{d}^k\end{bmatrix}^T + \begin{bmatrix}\vec{\alpha}_{a,1}\\ \vec{\alpha}_{a,2}\\ \vdots\\ \vec{\alpha}_{a,m}\end{bmatrix}\begin{bmatrix}\vec{a}\\ \vec{a}^2\\ \vdots\\ \vec{a}^m\end{bmatrix}^T + \begin{bmatrix}\alpha_g\\ \alpha_\ell\\ c\end{bmatrix}\begin{bmatrix}g\\ \ell\\ 1\end{bmatrix}^T \tag{2}$$

where

- v is the continuous average speed (i.e., the average speed without idling). We also consider the higher powers of v like $v^2, ..., v^r$.
- $\vec{d} = (\tau_d, \mu_d, \sigma_d)$ is the deceleration tuple:
  - τ_d is the total duration of deceleration.
  - μ_d is the mean deceleration (i.e., the sum of deceleration values divided by the deceleration duration).
  - σ_d is the standard deviation of deceleration.

  We denote the higher powers of components in the deceleration tuple by $\vec{d}^k = (\tau_d^k, \mu_d^k, \sigma_d^k)$.
- $\vec{a}$ is the acceleration tuple (similar to $\vec{d}$).
- g is the mean absolute value of gyroscope along the moving direction.
- ℓ is the auxiliary load of idling, which is the baseline measurement when the vehicle is not moving.

[1]Some of the variables are selected based on [6], which analyzed more than 20 thousand data points from 45 drivers to identify the most significant factors of fuel consumption and emission.

- c is a normalization constant.
- $\alpha_v \triangleq (\alpha_{v,1}, ..., \alpha_{v,r})$, $\alpha_d \triangleq (\alpha_{d,1}, ..., \alpha_{d,k})$, $\alpha_a \triangleq (\alpha_{a,1}, ..., \alpha_{a,k})$, α_g, α_ℓ are the corresponding coefficients.

4.2.2 Idle Vehicle Energy Consumption

Similarly, we rely on a blackbox approach to estimate the idle vehicle energy consumption. We drop the subscript $_{\mathsf{D,V,R}}$ for brevity. With respect to a particular combination of $(\mathsf{D}, \mathsf{V}, \mathsf{R})$, we estimate the idle vehicle energy consumption E^{id} (denoted by $\hat{E}^{\text{id}}$) by a linear equation:

$$\hat{E}^{\text{id}} = \beta\mu\ell \tag{3}$$

where

- μ is the total idle duration.
- ℓ is the auxiliary load of idling.
- β is a coefficient.

The auxiliary load of idling ℓ can provide the baseline energy consumption of an idle vehicle.

4.3 Estimation of Coefficients

The coefficients $(\alpha_v, \vec{\alpha}_d, \vec{\alpha}_a, \alpha_g, \alpha_\ell, c, \beta)$ in Eqns. (2)-(3) can be estimated by the standard regression method, if sufficient measured data $(v, \vec{d}, \vec{a}, g, \ell, \mu)$ and the respective energy consumption data $(\hat{E}^{\text{mv}}, \hat{E}^{\text{id}})$ are provided. We assume that each driver-vehicle pair (D, V) has collected sufficient historical personal driving data, and hence, the coefficients can be estimated for the respective environment R. One notable advantage of regression method is that it is less susceptible to random noise, which can arise from various sources (e.g., due to time synchronization in data sampling, mechanic damping, inaccurate measurements).

5. INTERPOLATING CROWD-SOURCED DATA

Given a dataset from crowd-sourced data collection, a data point can be visualized as a point in a 3-dimensional Euclidean space, indexed by $(\mathsf{D}, \mathsf{V}, \mathsf{R})$. The crowd-sourced dataset is usually sparse, consisting of a skewed and clustered distribution of data points. In order to predict the driving energy consumption for the data points that are not present in crowd-sourced data, we interpolate the missing data points to cover the space of dataset. An illustration is depicted in Fig. 2. We next describe the interpolation methods by comparison with the average and matrix factorization.

5.1 Comparison with the Average

The simplest approach for estimating driving energy consumption is to rely on the global characteristics in the dataset, for example, based on the average data values (e.g., average speed) from crowd-sourced data. However, each driver may deviate considerably from the average data values. To compensate for the deviations, incorporating a personalized adjustment can improve the prediction accuracy.

Let $x_{\mathsf{D,V,R}}$ be a measurement (e.g., v, $\vec{d}$ or $\vec{a}$) for tuple (D, V) in environment R, and the average data value be $\bar{x}_{\mathsf{R}}$. We use an adjustment function $\mathcal{D}^x_{\mathsf{D,V}}(\cdot)$ to convert the average data value to the personal data value by:

$$x_{\mathsf{D,V,R}} = \mathcal{D}^x_{\mathsf{D,V}}(\bar{x}_{\mathsf{R}}) \tag{4}$$

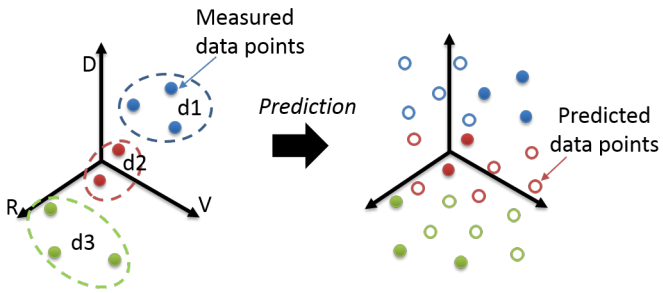

Figure 2: Measured and interpolated data points in a sparse dataset.

A simple adjustment function can be considered using regression model:

$$\mathcal{D}^x_{\mathsf{D,V}}(\bar{x}_{\mathsf{R}}) = \eta^1_{\mathsf{D,V}}\bar{x}^2_{\mathsf{R}} + \eta^2_{\mathsf{D,V}}\bar{x}_{\mathsf{R}} + \eta^3_{\mathsf{D,V}} \tag{5}$$

5.2 Matrix Factorization

Collaborative filtering is a domain-free approach, relying on the identification of abstract latent factors, which is based on the local characteristics among similar data points. Matrix factorization is a popular approach of constructing latent factors, which has been implemented in recommendation system [11] and other large-scale problems [7].

Consider an example of sparse matrix $X = [x_{\mathsf{D,V,R}}]$ of n pairs of (D, V) and m road segments R, as shown in Table 1, in which each entry represents a measurement (e.g., v, $\vec{d}$ or $\vec{a}$). Note that some data points may be missing in X (denoted by "?").

D,V \ R	1	2	3	4	...	m
1	67	74	?	32	...	50
2	54	?	83	44	...	65
⋮	?	74	53	?	...	?
n	?	66	58	?	...	88

Table 1: An example of sparse matrix X of vehicle speed v.

The basic idea of matrix factorization is to find two low-rank ($n \times k$ and $m \times k$) matrices, P and Q, such that PQ^T can approximate X. Namely,

$$X \approx PQ^T = \hat{X} \tag{6}$$

P and Q can be regarded as mappings to reduce the m, n-dimensional space of the original dataset to a k-dimensional space of latent factors, where $k \ll \min(m, n)$. We denote the entry at the i-th column and the j-th row of X be x_{ij}. Let $\mathcal{M}$ be the set of collected data points.

The objective of matrix factorization is find P, Q such that

$$\min_{P,Q} \sum_{i,j \in \mathcal{M}} (x_{ij} - p_i q_j^T)^2 + \lambda_P ||p_i||^2 + \lambda_Q ||q_j||^2 \tag{7}$$

where p_i is the i-th row vector of P, and q_j is the j-th column vector of Q. Since factorization may cause over-fitting, λ_P and λ_Q are used to regularize the fitting.

There are two popular approaches to compute P, Q in Eqn. (7): stochastic gradient descent [7] and alternating

least squares [11]. In this paper, we utilize stochastic gradient descent. The basic idea is to go through all x_{ij} in X. For each x_{ij}, determine the corresponding factor vectors p_i and q_j. Then, compute the approximate value by $p_i q_j^T$ and update the parameters by:

$$\begin{aligned} p_i &\leftarrow p_i + \epsilon(e_{ij} q_j - \lambda_P p_i) \\ q_j &\leftarrow q_j + \epsilon(e_{ij} p_i - \lambda_Q q_j) \end{aligned} \tag{8}$$

where $e_{ij} = x_{ij} - p_i q_j^T$ represents the difference between approximate value and actual value and ϵ is the learning rate. Once P, Q are determined, the estimation of a missing data $\hat{x}_{ij}$ can be estimated by $\hat{x}_{ij} = p_i q_j^T$. All measurements (e.g., v, $\vec{d}$ or $\vec{a}$) can be substituted and estimated using matrix factorization. The estimated values can be utilized in the driving energy consumption prediction.

6. PRIVACY-ENHANCING MECHANISMS

We describe mechanisms to enhance the privacy in the data collection process. Our approach is based on introducing random perturbation to obfuscate any inference about users' personal information. We assume a trusted data collector. However, the mechanisms can be adopted to consider an untrusted data collector, following the approach in [9].

6.1 Definition

Consider a dataset denoted by matrix $X = [x_{\mathsf{D},\mathsf{V},\mathsf{R}}]$ which represents a measurement (e.g., v, $\vec{d}$ or $\vec{a}$). A central principle of defining privacy is that one would not be able to infer the input dataset from the prediction output. Let $h(\cdot)$ be a randomized mapping (which incorporates random perturbation) from the input dataset to the prediction output, $\hat{X} = h(X)$. Specifically, given two input datasets X_1 and X_2, if they are not distinguishable under $h(\cdot)$ with an exponentially small probability, then it is unlikely to infer X_1 or X_2 from the prediction output, and hence, achieving a considerable extent of privacy, if X_1 and X_2 differentiate in a user's contributed data.

Formally, we consider an adjacency relation "$\approx$", which is a symmetric and transitive relation, such that $X_1 \approx X_2$, if X_1 and X_2 differentiate in certain contributed data from a user. The extract definition of "$\approx$" depends on the the level of privacy setting, which will be explained in the subsequent section. $h(\cdot)$ is said to satisfy *ϵ-differential privacy*, if

$$\mathbb{P}\big\{h(X_1) \in S\big\} \leq e^{\epsilon} \cdot \mathbb{P}\big\{h(X_2) \in S\big\} \tag{9}$$

for any $X_1 \approx X_2$, any range S of $h(\cdot)$ and an arbitrarily small constant ϵ.

6.2 Laplacian Mechanisms

A popular method to devise ϵ-differentially private mechanisms is based on random perturbation with Laplace probability distribution. Denote by $\texttt{Laplace}(\mu, b)$ a random variable, with probability density function $f(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$.

Let

$$d_{\max} \triangleq \max_{X_1 \approx X_2} \|h(X_1) - h(X_2)\| \tag{10}$$

and

$$h_\epsilon(X) \triangleq h(X) + \eta \tag{11}$$

where $\eta \sim \texttt{Laplace}(0, \frac{d_{\max}}{\epsilon})$ is a random variable with Laplace probability distribution. Then, it can be shown [1, 9] that $h_\epsilon(\cdot)$ satisfy *ϵ-differential privacy*.

Laplacian mechanisms have been proposed to achieve differential privacy considering diverse models of mapping $h(\cdot)$, such as linear classifiers, support vector machines, matrix factorization models.

6.3 Adjacency Relation

To properly apply Laplacian mechanisms, we consider the following settings of adjacency relation "$\approx$" to realize different level of privacy settings, and the respective $d_{\max}$.

- **Complete Privacy**: "$\approx$" is defined as X_1 and X_2 that differs in only one single entry. Let $X_1 = [x^1_{\mathsf{D},\mathsf{V},\mathsf{R}}]$ and $X_1 = [x^2_{\mathsf{D},\mathsf{V},\mathsf{R}}]$. Formally, we set $X_1 \approx X_2$, if and only if $x^1_{\mathsf{D}',\mathsf{V}',\mathsf{R}'} \neq x^2_{\mathsf{D}',\mathsf{V}',\mathsf{R}'}$ for only one tuple $(\mathsf{D}', \mathsf{V}', \mathsf{R}')$, otherwise, $x^1_{\mathsf{D},\mathsf{V},\mathsf{R}} = x^2_{\mathsf{D},\mathsf{V},\mathsf{R}}$.

  Let $[x_{\min}, x_{\max}]$ be the range of each entry $x_{\mathsf{D},\mathsf{V},\mathsf{R}}$ such that $x_{\mathsf{D},\mathsf{V},\mathsf{R}} \in [x_{\min}, x_{\max}]$. Note that when a measurement is absent (i.e., $x_{\mathsf{D},\mathsf{V},\mathsf{R}} =$ "?"), we assume that it takes any arbitrary default value in $[x_{\min}, x_{\max}]$. Therefore, $d_{\max} = |x_{\max} - x_{\min}|$.

- **Type-revealing Privacy**: "$\approx$" is defined as X_1 and X_2 that differs in only one single entry that belongs to the same driver type. Let $\mathbb{D}$ be a set of driver types, each $\mathcal{D} \in \mathbb{D}$ is a set of drivers, who share similar driving habits. Formally, $X_1 \approx X_2$, if and only if $x^1_{\mathsf{D}',\mathsf{V}',\mathsf{R}'} \neq x^2_{\mathsf{D}'',\mathsf{V}',\mathsf{R}'}$ for only one pair $(D', D'') \in \mathcal{D}$ for some $\mathcal{D} \in \mathbb{D}$, otherwise, $x^1_{\mathsf{D},\mathsf{V},\mathsf{R}} = x^2_{\mathsf{D},\mathsf{V},\mathsf{R}}$.

  Let $[x^{\mathcal{D}}_{\min}, x^{\mathcal{D}}_{\max}]$ be the type-specific range of each entry $x_{\mathsf{D},\mathsf{V},\mathsf{R}}$ such that $\mathsf{D} \in \mathcal{D}$. Note that when a measurement is absent (i.e., $x_{\mathsf{D},\mathsf{V},\mathsf{R}} =$ "?"), we assume that it takes any arbitrary default value in $[x^{\mathcal{D}}_{\min}, x^{\mathcal{D}}_{\max}]$. Therefore, $d^{\text{ty}}_{\max} = \max_{\mathcal{D} \in \mathbb{D}} |x^{\mathcal{D}}_{\max} - x^{\mathcal{D}}_{\min}|$. Note that $[x^{\mathcal{D}}_{\min}, x^{\mathcal{D}}_{\max}] \subseteq [x_{\min}, x_{\max}]$, and hence, $d^{\text{ty}}_{\max} \leq d_{\max}$.

- **Identity-revealing Privacy**: "$\approx$" is defined as X_1 and X_2 that differs in only one single entry of a particular driver and the predicted value without the driver. Let $\hat{x}_{\mathsf{D},\mathsf{V},\mathsf{R}}$ be the predicted value. Suppose that $\hat{x}_{\mathsf{D},\mathsf{V},\mathsf{R}} \in [x^{\mathsf{D},\mathsf{V}}_{\min}, x^{\mathsf{D},\mathsf{V}}_{\max}]$. When a measurement is absent (i.e., $x_{\mathsf{D},\mathsf{V},\mathsf{R}} =$ "?"), we assume that it takes any arbitrary default value in $[x^{\mathsf{D},\mathsf{V}}_{\min}, x^{\mathsf{D},\mathsf{V}}_{\max}]$. Therefore, $d^{\text{id}}_{\max} = \max_{\mathsf{D},\mathsf{V}} |x^{\mathsf{D},\mathsf{V}}_{\max} - x^{\mathsf{D},\mathsf{V}}_{\min}|$. Note that $[x^{\mathsf{D},\mathsf{V}}_{\min}, x^{\mathsf{D},\mathsf{V}}_{\max}] \subseteq [x^{\mathcal{D}}_{\min}, x^{\mathcal{D}}_{\max}]$, and hence, $d^{\text{id}}_{\max} \leq d^{\text{ty}}_{\max}$.

6.4 Estimation Error of Privacy Models

We implement the complete privacy based on Laplacian mechanism to the models.

1. Comparison with the Average:

   First we apply Laplacian mechanisms to comparison with the average model. Define the mapping by $h(X) = [\bar{x}_{\mathsf{R}}]$, where

   $$\bar{x}_{\mathsf{R}} = \sum_{\mathsf{D},\mathsf{V}} x_{\mathsf{D},\mathsf{V},\mathsf{R}} \tag{12}$$

   We suppose that the data collector will release $[\bar{x}_{\mathsf{R}}]$ to the public data repository. When $[\bar{x}_{\mathsf{R}}]$ is known, each driver can predict $\hat{x}_{\mathsf{D},\mathsf{V},\mathsf{R}}$ individually using Eqn. (5).

   We perturb the aggregate average speed information, and use it to create the personal adjustment functions

$\mathcal{D}^x_{\mathsf{D},\mathsf{V}}(\bar{x}_\mathsf{R})$. We evaluate the error between true personalized speed and estimated personalized speed using perturbed adjustment function.

2. Matrix Factorization:

   For matrix factorization, we assume that the data collector will first solve Eqn. (7) to obtain P, Q. Next, from the data collector we will solve $\tilde{Q}$ from the following problem given P:

$$\tilde{Q} = \arg\min_{Q} \sum_{i,j \in \mathcal{M}} (x_{ij} - p_i q_j^T)^2 + \lambda_Q ||q_j||^2 \qquad (13)$$

   The above problem can also be solved using stochastic gradient descent. Let $h(X) = \tilde{Q}$, and $\tilde{Q}$ will be released to the public data repository. When $\tilde{Q}$ is known, each driver can predict $\hat{x}_{\mathsf{D},\mathsf{V},\mathsf{R}}$ individually.

Fig. 3 shows the results of the error, we found that the estimation error drops below 0.15 when the value of privacy parameter ε is $\ln 3$ which is considered as the acceptable levels of privacy in the literatures. In addition, we found that the matrix factorization method is less affected by the perturbed features.

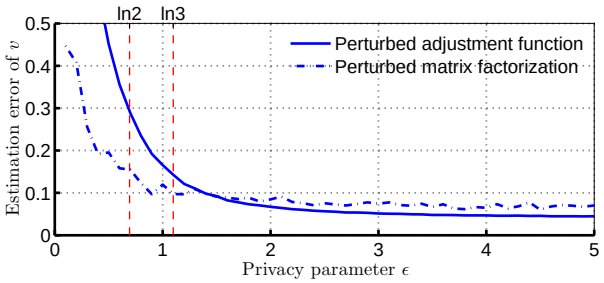

Figure 3: Estimation error of the features by the perturbed personalized models.

6.5 The Effect of Perturbed Features to Energy Consumption Estimation

We examine the predictions of energy consumption to ensure the perturbed features provide acceptable estimations. The perturbed features are utilized in Eqn. (2) and Eqn. (3) for the energy consumption predictions of the vehicles. Fig. 4 depicts the energy prediction error using perturbed features estimated by the matrix factorization method. We observe $ln3$- differential privacy provides below 10% energy estimation error.

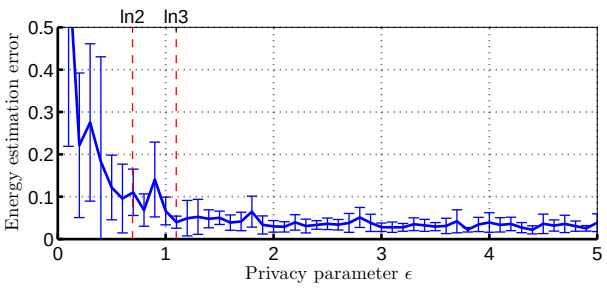

Figure 4: Energy estimation error of the perturbed matrix factorization.

7. CONCLUSION

We consider the privacy of participatory sensing data from the vehicles and discuss the estimation error of privacy-enhanced data used to predict the energy consumption of the vehicles. The privacy enhancing mechanism based on Laplacian mechanism is implemented to perturb the sensing data and the energy consumption estimation models. The estimation errors of the features using two perturbed models are below 10% with $ln3$-differential privacy, and we also observe that the sensing data of the same privacy level can achieve 10% prediction error of the energy prediction models. Future work will include evaluations of feature/energy prediction error using the type-revealing privacy and identity-revealing privacy mechanisms, furthermore, the best stage of privacy perturbation will be evaluated.

8. REFERENCES

[1] A. Berlioz, A. Friedman, M. A. Kaafar, R. Boreli, and S. Berkovsky. Applying differential privacy to matrix factorization. In *ACM Conf. on Recommender Systems*, 2015.

[2] K. Boriboonsomsin and M. J. Barth. Impacts of road grade on fuel consumption and carbon dioxide emissions evidenced by use of advanced navigation systems. *J. of the Transportation Research Board*, 2139:21–30, 2009.

[3] A. T. Campbell, S. B. Eisenman, N. D. Lane, E. Miluzzo, R. A. Peterson, H. Lu, X. Zheng, M. Musolesi, K. Fodor, and G.-S. Ahn. The rise of people-centric sensing. *IEEE Internet Computing*, 12(4):12–21, 2008.

[4] A. Cappiello, I. Chabini, E. K. Nam, A. Lue, and M. A. Zed. A statistical model of vehicle emissions and fuel consumption. In *IEEE ITSC*, 2002.

[5] S. Dornbush and A. Joshi. Streetsmart traffic: Discovering and disseminating automobile congestion using VANET. In *IEEE VTC*, 2007.

[6] E. Ericsson. Indepentent driving pattern factors and their influence on fuel-use and exhaust emission factor. *J. of Transportation Research*, 6(5):325–345, 2001.

[7] R. Gemulla, P. J. Haas, E. Nijkamp, and Y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *ACM SIGKDD*, 2011.

[8] S. Grubwinkler and M. Lienkamp. A modular and dynamic approach to predict the energy consumption of electric vehicles. In *Conf. on Future Automotive Technology*, 2013.

[9] J. Hua, C. Xia, and S. Zhong. Differentially private matrix factorization. In *IJCAI*, 2015.

[10] D. Karbowski, S. Pagerit, and A. Calkins. Energy consumption prediction of a vehicle along specified real-world trip. In *IEEE EVS*, 2012.

[11] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *J. of Computer*, 42:30–37, 2009.

[12] V. Nikolaenko, S. Ioannidis, U. Weinsberg, M. Joye, N. Taft, and D. Boneh. Privacy-preserving matrix factorization. In *ACM SIGSAC Conf. on Computer and Communications Security*, 2013.

[13] C.-M. Tseng and C.-K. Chau. Personalized prediction of driving energy consumption based on participatory sensing. Technical report, Masdar Institute, 2016.

[14] C.-M. Tseng, C.-K. Chau, S. Dsouza, and E. Wilhelm. A participatory sensing approach for personalized distance-to-empty prediction and green telematics. In *ACM E-energy*, 2015.

[15] C.-M. Tseng, S. Dsouza, and C.-K. Chau. A social approach for predicting distance-to-empty in vehicles. In *ACM E-energy*, 2014.

[16] E. Wilhelm, J. Siegel, S. Mayer, L. Sadamori, S. Dsouza, C.-K. Chau, and S. Sarma. CloudThink: A scalable secure platform for mirroring transportation systems in the cloud. *Transport*, 30(3), 2015.

[17] Q. Yang, K. Boriboonsomsin, and M. Barth. Arterial roadway energy/emissions estimation using modal-based trajectory reconstruction. In *IEEE ITSC*, 2011.