Privacy-Preserving Continuous Data Collection for Predictive Maintenance in Vehicular Fog-Cloud

Qinglei Kong, *Student Member, IEEE*, Rongxing Lu, *Senior Member, IEEE*, Feng Yin, *Member, IEEE*, and Shuguang Cui, *Fellow, IEEE*

***Abstract*— With the advances of Internet of Things (IoT) solutions in intelligent transportation systems, collected vehicle data can produce insights on emerging vehicular phenomenon, and further contribute to the further improvement of innovative and efficient vehicular systems. Particularly, by leveraging data collected from vehicle sensors and maintenance models constructed from operation and repair history, predictive maintenance aims to detect the anomalies of vehicles and provide early warnings before the occurrence of failure. However, privacy preservation still remains as one of the top concerns for vehicle owners in predictive maintenance, as the sensory data could potentially violate their location and identity privacy. To address this challenge, in this article, we propose a privacy-preserving and verifiable continuous data collection scheme with the intent of predictive maintenance in vehicular fog, which gathers and organizes the sensor data of each individual vehicle on a sliding window basis. Specifically, our proposed scheme exploits the homomorphic Paillier cryptosystem and truncated α-geometric technique to protect the content of each individual piece of sensory data. Meanwhile, our proposed scheme also aggregates and authenticates the collected sensory data reports on a time-series sliding window basis, which achieves the continuous observation of the recently collected vehicular sensory data. Detailed security analysis is carried out to demonstrate the security properties of our proposed scheme, including confidentiality, authentication and privacy preservation. In performance evaluations, we also compare our proposed scheme with a traditional scheme, and our scheme shows great improvement in terms of communication and computation overheads. Furthermore, to show the feasibility of our proposed scheme, we also compare and discuss the expected squared error introduced by the differential privacy mechanism.**

***Index Terms*— Predictive maintenance, privacy preservation, vehicular fog-cloud.**

Manuscript received March 26, 2020; revised May 25, 2020; accepted July 9, 2020. This work was supported in part by the Key Area Research and Development Program of Guangdong Province under Grant 2018B030338001, in part by the National Key Research and Development Program of China under Grant 2018YFB1800800, in part by the Natural Science Foundation of China under Grant NSFC-61629101, in part by the Guangdong Research Project under Grant 2017ZT07X152, and in part by the Shenzhen Key Lab Fund under Grant ZDSYS201707251409055. The Associate Editor for this article was S. Garg. *(Corresponding author: Qinglei Kong.)*

Qinglei Kong is with the Future Network of Intelligence Institute (FNii), The Chinese University of Hong Kong, Shenzhen 518172, China, and also with the University of Science and Technology of China, Hefei 230052, China (e-mail: kongqinglei@cuhk.edu.cn).

Rongxing Lu is with the Faculty of Computer Science, University of New Brunswick, Fredericton, NB E3B 5A3, Canada (e-mail: rlu1@unb.ca).

Feng Yin and Shuguang Cui are with the Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen 518172, China, and also with the Future Network of Intelligence Institute (FNii), The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: yinfeng@cuhk.edu.cn; shuguangcui@cuhk.edu.cn).

Digital Object Identifier 10.1109/TITS.2020.3011931

I. Introduction

NOWADAYS, by harnessing the power of Internet of Things (IoT) [1]–[3] and predictive analytics in intelligent transportation systems, predictive maintenance is widely recognized for providing operational insights and warnings of vehicles [4]–[6]. Instead of merely reflecting the current engine status, the rich body of historical vehicular operation and repair data can also build out the entire automotive condition landscape and further construct a superb predictive model with artificial intelligence, that spots emerging trends before they turn into system failure [4], [5]. For example, a predictive maintenance model is constructed from Volvo service records and the logged on-board data, covering parameters like mileage, engine hours and fuel consumption, which achieves the prediction for air compressor repair of commercial trucks and buses [7]. By exploiting various vehicle attributes like mileage, age, and vehicle type, a predictive maintenance scheme is proposed in [8], which can effectively reduce the time and labor associated with inspections and repairs of vehicle fleets.

Nevertheless, before the wide deployment of predictive maintenance in vehicular fog, issues like privacy preservation need to be deliberatively considered and addressed [9]. Research in [10] studies the possibility of identifying one driver from snippets of sensory data. Experimental results show that data streams may capture driver actions, and further lead to the potential privacy leakage. In intelligent transportation systems, statistics collected from on-board devices and sensors are highly location-dependent, and the sensory data will pose serious privacy threats, especially in terms of movement trajectory leakage and personal preference disclosure. To achieve privacy preserving of data collection, various security schemes have been broadly studied within the industrial IoT framework. In [11], a privacy-preserving multi-dimensional sensory data aggregation of multiple users in smart grid is achieved with the homomorphic encryption techniques. Under the vehicular sensing settings, privacy preserving data collection schemes are also proposed in [12], [13], which achieve aggregated sensory data collection from a certain coverage area during a given time period. Specifically, all of the above schemes preserve the privacy of an individual user through multiple users' data aggregation, and the aggregation results can be used to describe a phenomenon in the spatial-temporal domain. However, the above schemes cannot be applied to predictive maintenance in intelligent transportation systems,

as it requires constant observation from the perspective of each individual vehicle.

Differential privacy brings a limitation towards privacy loss, which constrains the effect brought by each individual piece of data towards the final computation result [14]–[16]. To achieve fault tolerant data aggregation, a lightweight privacy-preserving aggregation scheme is proposed in [17] by combining homomorphic encryption and noise extracted from geometric distribution. A collusion-resistant and differentially-private aggregation scheme for star networks is proposed in [18], which exploits the homomorphic encryption for data aggregation, under the semi-honest participants setup. However, the above schemes cannot be directly adopted towards the continuous query for the time-series data processing, as they only consider the privacy loss in the discrete query process. An expiration mechanism taking privacy budget into consideration is proposed in [19], which manages the floating car data (FCD) records lifetime for intelligent transportation system. In addition, the authors in [20] proposed a differentially private sliding-window count querying scheme for a data stream management system (DSMS), which derives the statistical information of data contained in a window. In our proposed scheme, as the predictive maintenance process requires a certain level of data accuracy, adding noise towards sensory dat, especially in the context of the sequential and continuous sensory data gathering.

To address the above-mentioned challenges, in this article, we propose a privacy-preserving and verifiable continuous sensory data collection scheme in vehicular fog. Specifically, the contributions of this article are three-fold:

1) First, we devise a privacy-preserving continuous vehicular sensory data generation and aggregation scheme for the intent of predictive maintenance. Specifically, each fog device, i.e., service and maintenance store, treats the sensory data collected during each time slot as a module, assembles the modules into a time-series sliding window, and periodically uploads the aggregation result towards the cloud owner with a predictive digital model. As our proposed scheme involves the continuous observation of correlated sliding windows, we also exploit the truncated α-geometric mechanism to achieve ϵ-differential privacy. In addition, each individual piece of sensory data can still be recovered with the collaboration of the vehicle and cloud server when necessary.
2) Second, we exploit an identity-based signature scheme to authenticate the origins and verify the correctness of the sliding window aggregation results with noise injected. In our proposed scheme, even though the value space is constraint, the content of each individual data piece cannot be recovered from signatures by repetitively exhausting all the possible data values.
3) Finally, detailed security analysis is conducted to validate the security properties, i.e., confidentiality, verifiability and privacy preservation. In performance evaluations, we compare the performance of the proposed scheme with a traditional scheme, which adds noise towards the collected sensory data in each time slot. We compare the introduced overheads, in terms of computation complexities and communication costs, which also corroborates the applicability of our proposed scheme. Furthermore, we explore a utility metric to show the feasibility of both schemes, which is denoted as the expected squared error.

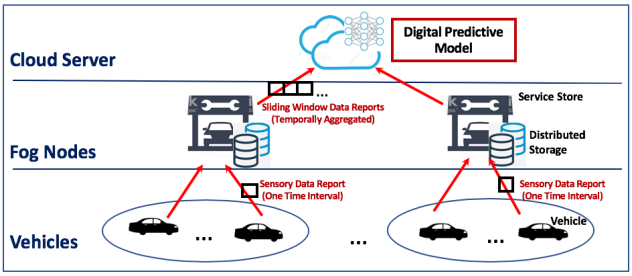


Fig. 1. Proposed continuous temporal data aggregation system.

The remainder of this article is structured as following. We introduce our system model, present our security requirements, and identify our design goals in Section II, and show the preliminaries in Section III. Then we present our proposed privacy preserving data filtering scheme under the embezzlement investigation setting in Section IV. The security analysis and performance evaluations are shown in Sections V and VI, respectively. Related work is presented in Section VII, and we conclude the paper in Section VIII.

II. System Model, Security Requirements, and Design Goals

In this section, we introduce the system model, present the security requirements, and identify the design goals.

A. System Model

In our system model, we consider a sensory data-driven predictive maintenance framework in vehicular fog-cloud, which potentially reduces engine downtime and provide early warnings in intelligent transportation systems. Since predictive maintenance emphasizes on data collected during a few recent time intervals, we consider the sliding window based vehicle sensor data collection and aggregation paradigm. As shown in Fig. 1, the proposed system consists of three types of entities: vehicles, fog nodes, and a cloud server, whose functionalities are shown as follows.

- *Vehicles.* Each vehicle periodically collects a plethora of vehicular sensory data indicating its current operational condition and auto parts functionality, such as trip distance, relative distance, axle angle, oil temperature/pressure, etc. As an end node, each vehicle organizes and uploads the sensory data reports towards the corresponding fog node, i.e., the contracted service and maintenance store. Note that due to the dynamically moving characteristic of vehicles, our proposed scheme exploits the service and maintenance stores as fog nodes to store and organize sensory data reports, instead of exploiting base stations as fog nodes.

- *Fog Nodes.* Each store plays the role of a fog computing and storage node, and it is responsible for the maintenance and repair of contracted customer vehicles. To be more specific, it receives the vehicular sensory data, structures the data on a sliding window basis, and delivers the organized data towards the cloud server. In addition, the service store is also a fog storage device, which keeps the received sensory data at its local storage and waits to be exploited for diagnosis when necessary.
- *Cloud Server.* The cloud server owns a digital model representing the status of auto components belonging to a few vehicle types. The model could be built with simulating what-if scenarios, actual field testing, and repair observation, which helps to determine and predict the health of auto parts. By exploiting the sliding window aggregation results of the current and historical sensory data, the digital predictive model can predict the occurrence of system breakdowns, and transmit warnings towards the corresponding fog node and vehicles.

Communication Model: The connections between vehicles and fog nodes are realized through base stations via wireless links like 5G connection, and those between fog nodes and the cloud server, are realized through either wired link or any other link with high bandwidth and low transmission delay, for instance, optical fiber cables.

B. Security Requirements

In the threat model, we first assume the cloud server is honest-but-curious. That is, it will follow the defined protocol, but it will try to identify the content of an individual data report. For each fog node, it also assumes to be honest-but-curious. Specifically, it will follow the defined protocol, but it will also try to infer each individual data report content and learn the aggregation result. Furthermore, we assume there is no collusion between any two entities in our proposed scheme. In addition, we also assume there exists an active adversary to eavesdrop and modify data reports during data transmission. Thus, to achieve privacy-preserving sensory data collection for predictive maintenance, the following security requirements must be met.

- *Confidentiality.* During the data transmission process, to protect the content of each individual piece of sensory data, the collected sensory data report should be protected. Meanwhile, only the cloud server can derive the sliding window aggregation result. In addition, during sensory data aggregation process at fog node side, the fog node cannot obtain the aggregated result.
- *Verifiability.* In the proposed scheme, as there exists an active adversary, the correctness of the results should be guaranteed. That is, after the sliding window aggregation result recovery, the cloud server should authenticate the origins and verify its correctness. Moreover, as the cloud server may connect to multiple fog nodes, it may receive a huge number of sensory data reports, and the proposed scheme should support batch authentication. In addition, when an individual data report needs to be recovered for the maintenance purpose, the proposed scheme also be able to authenticate each individual piece of sensory data.
- *Privacy Preservation.* As sensory data are highly individual- and location-dependent, the cloud server should not learn the content of any individual piece of sensory data, and it should only learn each vehicle's sliding window aggregation result. Meanwhile, given two adjacent sliding window aggregation results, whose inputs are only differing in one pair of sensory data reports, the knowledge extracted from these two reports should be limited. In addition, as the possible value space is limited, based on the signature of each data report, neither the fog node nor the cloud server can infer the data report content.

C. Design Goals

Under the aforementioned system model and security requirements, our design goal is to create a privacy-preserving sliding window data collection scheme for predictive maintenance in vehicular fog. Specifically, the proposed scheme should achieve the following design goals.

The proposed scheme should achieve the above-mentioned security requirements. If the proposed scheme does not take security requirements into consideration, the individual piece of sensory data could be disclosed, which may further violate the vehicle's location privacy, and the correctness of data transmission may not be guaranteed. Then the vehicle owners may not be willing to get involved in the predictive maintenance process, and the status of the vehicle may not be effectively captured.

The proposed scheme should achieve the goal of flexibility. Even though the proposed scheme aims to derive the sliding window aggregation result, the content of each individual sensory data should also be recovered when necessary. Meanwhile, at the fog node side, to accurately identify a vehicle's problem during vehicle diagnosis. It should also reveal an individual sensory data towards the fog node upon the permission of the vehicle owner.

The proposed scheme should achieve the goal of high efficiency in terms of computational complexity and communication overhead and high utility in terms of accuracy. Although the cloud server, fog nodes, and vehicles possess strong computational ability, the introduced computational complexity should be deliberatively evaluated, especially when there involve a huge number of vehicles simultaneously sending sensory data reports together. Meanwhile, the proposed scheme should also take the communication overhead into consideration, especially the vehicle-to-fog overhead during the data collection process. In addition, as the proposed scheme is designed for the predictive maintenance application, the accuracy of the scheme should also be deliberatively evaluated.

III. Preliminaries

In this section, we briefly review the security techniques of bilinear maps, Paillier cryptosystem, ϵ-differential privacy and truncated α-geometric mechanism, which serve as the foundations of the proposed scheme.

A. Bilinear Maps

Given a security parameter κ_1, let $\mathbb{G}$ and $\mathbb{G}_T$ be two cyclic groups of the same large prime order $|q| = \kappa_1$.

1) Bilinearity: Given $\forall P, Q \in \mathbb{G}$, and $\forall a, b \in \mathbb{Z}_q$, we can derive $e(aP, bQ) = e(P, Q)^{ab}$;
2) Non-degeneracy: There exists $P, Q \in \mathbb{G}$, which satisfies the condition that $e(P, Q) \neq 1_{\mathbb{G}_T}$.
3) Computability: $\forall P, Q \in \mathbb{G}$, there exists an efficient algorithm to compute $e(P, Q)$.

Definition 1: A bilinear parameter generator $\mathcal{G}en$ denotes a probabilistic algorithm that takes a parameter κ_1 as input, and generates a 5-tuple $(q, \mathbb{G}, \mathbb{G}_T, e, P)$ as the output, where q is a prime number with $|q| = \kappa_1$, $\mathbb{G}$ is an additive cyclic group and $\mathbb{G}_T$ is a multiplicative cyclic group, $P \in \mathbb{G}$ is a generator, and $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ is non-degenerated and computable bilinear map.

B. Paillier Cryptosystem

To achieve privacy-preserving sensory data analysis, Paillier cryptosystem is exploited for achieving the homomorphic properties [21], which is utilized to protect the content of each individual piece of sensory data, and to achieve the secure aggregation of multiple data report. Specifically, Paillier cryptosystem consists of three algorithms: key generation, encryption and decryption.

- *Key generation.* Given a security parameter κ, two large prime numbers p_1 and q_1 are first chosen, where $|p_1| = |q_1| = \kappa$. Then the RSA modulus $n = p_1 \cdot q_1$ and the least common multiple $\lambda = lcm(p_1 - 1, q_1 - 1)$ are computed. Given a function $L(u) = \frac{u-1}{n}$, after choosing a generator $g \in \mathbb{Z}^*_{n^2}$, the value $\mu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$ is further calculated. Finally, the public key is $pk = (n, g)$, and the corresponding private key is $sk = (\lambda, \mu)$.
- *Encryption.* Given a message $m \in \mathbb{Z}_n$, choose a random number $r \in \mathbb{Z}^*_n$, and the ciphertext can be computed as $c = E(m) = g^m \cdot r^n \bmod n^2$.
- *Decryption.* Given a ciphertext $c \in \mathbb{Z}^*_{n^2}$, the corresponding plaintext can be recovered as $m = D(c) = L(c^\lambda \bmod n^2) \cdot \mu \bmod n$.

C. Differential Privacy

Differential privacy technique has received considerable attention in privacy-preserving statistical analysis. Intuitively, the core idea of a differential privacy mechanism is to add reasonable noise, such that the outputs are indistinguishable whenever an arbitrary item changes, whose core idea is summarized as follows [20].

Definition 2: (Neighboring Datasets). Datasets DS_1 and DS_2 are neighboring if they differ only at one data item, i.e., $DS_1 = (d_i, d_{i+1}, \ldots, d_{i+w-1})$ and $DS_2 = (d_{i+1}, d_{i+2}, \ldots, d_{i+w})$, where $d_j, j \in \{i, i+1, \ldots, i+w\}$ denotes the data generated at time j.

Differential privacy ensures that an adversary cannot distinguish two neighboring datasets based on their query results. Given a function $f(\cdot)$, a noisy mechanism $A(\cdot)$ for $f(\cdot)$ is a probabilistic function, gives an output $A(DS)$ with a certain distribution relying on the true output $f(DS)$. Note that $A(DS)$ may coincide with $f(DS)$, and we exploit $Pr[A(DS) = k]$ to denote the probability that $A(\cdot)$ applied to DS reports the answer k. Now we define how does a mechanism $A(\cdot)$ achieve ϵ-differential privacy.

Definition 3: A privacy mechanism $A(\cdot)$ gives ϵ-differential privacy, where ϵ is a non-negative real number reflecting the level of privacy, if for any two neighboring datasets DS_1 and DS_2 differing on at most one data item, and for every output $A(DS) = k$, it satisfies the condition:

$$Pr[A(DS_1) = k] \leq e^\epsilon \cdot Pr[A(DS_2) = k], \tag{1}$$

in which a lower ϵ guarantees a stronger privacy level with a higher perturbation noise. Meanwhile, the L_1-sensitivity of function $f(\cdot)$ is defined as

Definition 4: The sensitivity of a function $f(\cdot)$ is defined as

$$\Delta f = \max_{DS_1, DS_2} ||f(DS_1) - f(DS_2)||_1 \tag{2}$$

where DS_1 and DS_2 are the neighboring dataset with at most one different data item.

D. Truncated α-Geometric Mechanism

As our proposed scheme only supports integer values, we exploit a discrete version of Laplacian noise with fixed output range, for the noise generation. For a given function $f(\cdot)$ and a parameter value $\alpha \in (0, 1)$, the truncated α-geometric mechanism $Geom(\alpha)$ is an oblivious mechanism with an output range $M = \{0, 1, \ldots, m\}$. When the true function output is $f(DS)$, the mechanism has an output $A(DS) = f(DS) + r$ with range M, where r is a randomly added noise with the following distribution,

$$\begin{cases} Pr[r = -f(DS)] = \dfrac{\alpha^{f(DS)}}{1+\alpha} \\ Pr[r = \delta] = \dfrac{1-\alpha}{1+\alpha} \cdot \alpha^{|\delta|}, \\ \quad \text{where} - f(DS) < \delta < m - f(DS) \\ Pr[r = m - f(DS)] = \dfrac{\alpha^{m-f(DS)}}{1+\alpha}. \end{cases} \tag{3}$$

Meanwhile, this mechanism can be viewed as a discretized version of a continuous mechanism which adds random noise extracted from a Laplace distribution, where $\epsilon = \ln \frac{1}{\alpha}$.

IV. Proposed Privacy-Preserving Data Collection and Aggregation Scheme

In this section, we present the privacy-preserving sensory data collection and aggregation scheme in vehicular fog, which enables the continuous observation of data reports on a sliding window basis. We first describe the system initialization phase, then we show the data collection and aggregation phases, respectively.

A. System Initialization

We assume there exists a trusted authority (TA), i.e., automotive management authority, will bootstrap the entire system. Given a security parameter κ_1, TA generates the bilinear parameters by running $\mathcal{G}en(\kappa_1)$, and derives a 5-tuple denoted as $\{q, \mathbb{G}, \mathbb{G}_T, e, P\}$. Meanwhile, TA selects two hash functions, which are $H_1(\cdot) : \{0,1\}^* \rightarrow \mathbb{G}$ and $H_2(\cdot) : \{0,1\}^* \rightarrow \mathbb{Z}_q^*$, respectively. In addition, TA selects a random number $s \in \mathbb{Z}_q^*$ as a system secret key, and takes $Q = s \cdot P$ as the system public key. Finally, TA publishes the system parameters $params_s = \{q, \mathbb{G}, \mathbb{G}_T, e, P, Q, H_1, H_2\}$.

Given a security parameter κ, the cloud server initializes the Paillier cryptosystem, which the public and private key pair are denoted as $pk = (n, g)$ and $sk = (\lambda, \mu)$, respectively. Then the cloud server selects two prime numbers α and β, such that $\alpha^{u+1} < n$ and $w \cdot \beta < \alpha$, where u denotes sensory data dimension and w is the sliding window length. Specifically, during each time slot, each fog node organizes the u-dimension data records collected by each vehicle during the recent w time slots. In addition, the cloud server identifies a starting time point T_0, and defines a time slot length t, such that the sensory data is processed on a time slot basis. Finally, the cloud server publishes its system parameters, which is denoted as $params_m = \{(n, g), \alpha, \beta, u, w, T_0, t\}$.

During the registration of a vehicle with identity id_v, TA generates and delivers an identity-based secret key $sk_v = (P_{v,0}, P_{v,1})$, where $P_{v,j} = s \cdot H_1(id_v||j) \in \mathbb{G}$, $j \in \{0,1\}$. Then vehicle v also selects a secret value k_v and keeps it in the local storage.

B. Sensory Data Collection

For vehicle id_v, the u-dimension sensory data collected during the i-th time slot is denoted as $(d_{i,1}, d_{i,2}, \ldots, d_{i,u})$, which satisfies the condition that $d_j^{min} \leq d_{i,j} \leq d_j^{max}$, $j \in \{1, 2, \ldots, u\}$ and $d_j^{max} \cdot w \leq \beta$. Then vehicle v performs the following steps to generate a data report, which is

- Vehicle id_v first computes the value $k_{i,j} = H_2(k_v||i||j)$ for each data dimension. Meanwhile, vehicle id_v generates the data sequence $(e_{i,1}, e_{i,2}, \ldots, e_{i,u})$, in which $e_{i,j} = (d_{i,j} + k_{i,j}) \bmod \beta$. In addition, vehicle id_v structures the derived data sequence with a prime number α, and obtains a structured sensory data, which is

$$data_i = \sum_{j=1}^{u} \alpha^j \cdot e_{i,j}. \tag{4}$$

- In order to prevent the cloud server from learning the sensory data based on two consecutive sliding window aggregations, the noise extracted from the truncated α-geometric mechanism is added. Specifically, vehicle id_v selects a random number $r_{i,j}$ from the truncated α-geometric mechanism with the distribution $Geom(exp(-\frac{\epsilon}{\Delta d_j}))$ within a given value range $\{w \cdot d_j^{min}, w \cdot d_j^{min} + 1, \ldots, w \cdot d_j^{max}\}$, where $\Delta d_j = d_j^{max} - d_j^{min}$. To recover the sliding window aggregation result, vehicle v computes a value $k'_{i,j} = \sum_{o=i-w+1}^{i} -H_2(k_v||o||j)$ for each data dimension $j \in \{1, 2, \ldots, u\}$. Meanwhile, it generates another value sequence $(e'_{i,1}, e'_{i,2}, \ldots, e'_{i,w})$, in which $e'_{i,j} = (r_{i,j} + k'_{i,j}) \bmod \beta$. In addition, it structures the newly generated value sequence with a prime number α, and obtains the value r_i

$$r_i = \sum_{j=1}^{u} \alpha^j \cdot e'_{i,j}. \tag{5}$$

- Vehicle id_v encrypts $data_i$ and r_i with the cloud server's Paillier cryptosystem public key (n, g), and derives the ciphertext pair

$$\begin{cases} c_i = g^{data_i} \cdot s_{i,1}^n \mod n^2, \\ \hat{c}_i = g^{r_i} \cdot s_{i,2}^n \mod n^2, \end{cases} \tag{6}$$

where $(s_{i,1}, s_{i,2}) \in \mathbb{Z}_n^*$ are two random numbers selected by vehicle v.

- Vehicle id_v also selects two random numbers $(t_{i,1}, t_{i,2}) \in \mathbb{Z}_q^*$, and generates two pairs of identity-based signatures [22] with the collected sensory data sequence $(d_{i,1}, d_{i,2}, \ldots, d_{i,u})$ and the extracted noise sequence $(r_{i,1}, r_{i,2}, \ldots, r_{i,u})$, which are

$$\begin{cases} \sigma_{i,1} = t_{i,1} \cdot P, \\ \sigma_{i,2} = t_{i,1} \cdot H_1(id_s) + P_{v,0} \\ \qquad + (\sum_{j=1}^{u} d_{i,j} + H_2(k_v||i)) \cdot P_{v,1}, \end{cases} \tag{7}$$

where id_s is the identity of the fog node, i.e., the bound service provider of vehicle id_v, and $(\sigma_{i,1}, \sigma_{i,2})$ denotes the signature of sensory data collected during the i-th time slot.

$$\begin{cases} \hat{\sigma}_{i,1} = t_{i,2} \cdot P, \\ \hat{\sigma}_{i,2} = t_{i,2} \cdot H_1(id_s) + P_{v,0} \\ \qquad + (\sum_{j=1}^{u} r_{i,j} - \sum_{o=i-w+1}^{i} H_2(k_v||o)) \cdot P_{v,1}, \end{cases} \tag{8}$$

where $(\hat{\sigma}_{i,1}, \hat{\sigma}_{i,2})$ denotes the signature of the added noise during the i-th time slot.

Finally, vehicle v formulates a sensory data report $Report_i = id_v||c_i||\hat{c}_i||\sigma_{i,1}||\sigma_{i,2}||\hat{\sigma}_{i,1}||\hat{\sigma}_{i,2}||i$, and delivers $Report_i$ towards the fog node id_s.

C. Sliding Window Data Aggregation

During the i-th time slot, after receiving the ciphertext pair $(c_i, \hat{c}_i)$ from vehicle id_v, the fog node id_s organizes the involved encrypted sensory data reports $(c_{i-w+1}, c_{i-w+2}, \ldots, c_i, \hat{c}_i)$ for sliding window aggregation. Then, fog node id_s performs the following ciphertext aggregation process, which is

$$\begin{aligned} C_i^a &= (\prod_{o=i-w+1}^{i} c_o) \cdot \hat{c}_i \bmod n^2 \\ &= g^{(\sum_{o=i-w+1}^{i} data_i) + r_i} \cdot (\prod_{o=i-w+1}^{i} s_{o,1} \cdot s_{i,2})^n \bmod n^2. \end{aligned} \tag{9}$$

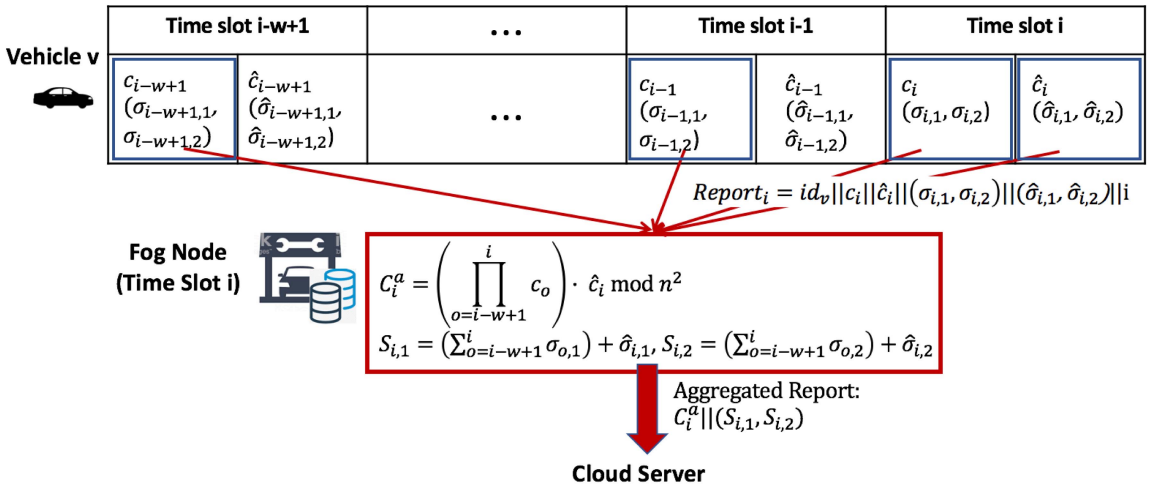


Fig. 2. Sensory data collection and aggregation at different time slots.

Then fog node id_s sends the ciphertext C_i^a towards the cloud server. Meanwhile, vehicle id_v organizes the signatures $((\sigma_{i-w+1,1}, \sigma_{i-w+1,2}), \ldots, (\sigma_{i,1}, \sigma_{i,2}), (\hat{\sigma}_{i,1}, \hat{\sigma}_{i,2}))$ with the following process, which is

$$\begin{cases} S_{i,1} &= \sum_{o=i-w+1}^{i} \sigma_{o,1} + \hat{\sigma}_{i,1} \\ &= (\sum_{o=i-w+1}^{i} t_{o,1} + t_{i,2}) \cdot P \\ S_{i,2} &= \sum_{o=i-w+1}^{i} \sigma_{o,2} + \hat{\sigma}_{i,2} \\ &= (\sum_{o=i-w+1}^{i} t_{o,1} + t_{i,2}) \cdot H_2(id_s) + (w+1) \cdot P_{v,0} \\ &\quad + \sum_{j=1}^{u} (\sum_{o=i-w+1}^{i} d_{o,j} + r_{i,j}) \cdot P_{v,1}, \end{cases} \tag{10}$$

where $(S_{i,1}, S_{i,2})$ is the aggregated signature pair of time slot i. In addition, the fog node sends the aggregated report $C_i^a||(S_{i,1}, S_{i,2})$ of time slot i towards the cloud server, as shown in Fig. 2.

For the cloud server, after receiving the aggregated ciphertext C_i^a, it performs the decryption process with its private key (λ, μ) and further obtains the value $data_i^a = \sum_{j=1}^{u} \alpha^j \cdot ((\sum_{o=i-w+1}^{i} d_{o,j}) + r_{i,j})$. Then it recovers the sliding window aggregation result for each data dimension, which is $data_{i,j}^a = (\sum_{o=i-w+1}^{i} d_{o,j}) + r_{i,j} \bmod \beta$, $j \in \{1, 2, \ldots, u\}$. With each aggregation result $data_{i,j}^a$, the cloud server authenticates the correctness of the recovered result, which is

$$\begin{aligned} e(S_{i,2}, P) \stackrel{?}{=} & \; e(S_{i,1}, H_1(id_s)) \\ & \cdot e((w+1) \cdot H_1(id_v||0), Q) \\ & \cdot e((\sum_{j=1}^{w} data_{i,j}^a) \cdot H_1(id_v||1), Q). \end{aligned} \tag{11}$$

If Eq. (11) is verified to be correct, the cloud server takes $(data_{i,1}^a, data_{i,2}^a, \ldots, data_{i,u}^a)$ as an input for the predictive maintenance process.

V. Security Analysis

In this section, we discuss the security properties of the proposed sliding window based data collection scheme. Based on the security requirements defined in Section II-B, the security properties are illustrated in terms of confidentiality, privacy preservation, and verifiability, respectively.

A. The Proposed Scheme Can Achieve the Security Goal of Confidentiality

Firstly, each individual piece of sensory data/ added noise is encrypted with the public key of the cloud server (n, g), such that any other entity cannot decrypt the corresponding ciphertexts (c_i, c_i'). As the exploited Paillier cryptosystem is proven to be semantically secure under the chosen plaintext attack, the individual piece of sensory data can be protected in the proposed scheme. Secondly, as the ciphertext of the homomorphic Paillier cryptosystem supports homomorphic addition, the fog node can derive the ciphertext of the sliding window based on the sensory data aggregation result without disclosing its content. Therefore, the security goal of confidentiality can be achieved in the proposed scheme.

B. The Proposed Scheme Can Achieve the Security Goal of Verifiability

Firstly, each individual piece of sensory data/ added noise is signed with an identity-based signature scheme, and then the signature pairs $(\sigma_{i,1}, \sigma_{i,2})$ and $(\hat{\sigma}_{i,1}, \hat{\sigma}_{i,2})$ are derived. Meanwhile, the exploited signature scheme is verified to be provably secure under Computational Diffie-Hellman against adaptive chosen identities and messages [22]. Secondly, the aggregation of the sliding window sequence signatures $((\sigma_{i-w+1,1}, \sigma_{i-w+1,2}), \ldots, (\sigma_{i,1}, \sigma_{i,2}))$ and the added noise signature $(\hat{\sigma}_{i,1}, \hat{\sigma}_{i,2})$ can also achieve the verification of the aggregated sliding window. In addition, since the exploited signature scheme also supports the batch authentication of signatures generated by different users, the proposed scheme can also achieve batch authentication of multiple vehicles' signatures. Thus, the security goal of verifiability can be achieved in the proposed scheme.

C. The Proposed Scheme Can Achieve the Security Goal of Privacy Preservation

Even though each individual piece of sensory data is encrypted with the public key of the Paillier cryptosystem (n, g), given the ciphertext pair (c_i, c_i'), the cloud server still cannot recover its true value and the added noise $(data_i, r_i)$

without learning the secret key k_v of vehicle v. To derive the sliding window sensory data aggregation result, the fog node aggregates the sequence $(c_{i-w+1}, c_{i-w+2}, \ldots, c_i)$ and the added noise $\hat{c}_i$ of time slot i. After decrypting the aggregated ciphertext C_i^a with its private key (λ, μ), the cloud server can obtain the aggregated sliding window sensory data with added noise $\sum_{o=i-w+1}^{i} d_o + r_i$, and further recover the aggregated sensory data in each data dimension $\sum_{o=i-w+1}^{i} d_{o,j} + r_{i,j}$.

In the sequel, we show that the sliding window aggregation value $data_i$ can achieve ϵ-differential privacy. Given two continuous temporally aggregated value $data_{i-1,j} = (\sum_{o=i-w}^{i-1} d_{o,j}) + r_{i-1,j}$ and $data_{i,j} = (\sum_{o=i-w+1}^{i} d_{o,j}) + r_{i,j}$, where $r_{i-1,j}$ and $r_{i,j}$ are two truncated α-geometric noise terms. For any integer $0 < k < n$, we can derive

$$\begin{aligned}\eta &= \frac{Pr(data_{i-1,j} = k)}{Pr(data_{i,j} = k)} \\ &= \frac{Pr(r_{i-1,j} = k - \sum_{o=i-w}^{i-1} d_{o,j})}{Pr(r_{i,j} = k - \sum_{o=i-w+1}^{i} d_{o,j})} \\ &= \alpha^{|k-\sum_{o=i-w}^{i-1} d_{o,j}| - |k-\sum_{o=i-w+1}^{i} d_{o,j}|}. \end{aligned} \tag{12}$$

Since $-|d_{i,j} - d_{i-w,j}| \leq |k - \sum_{o=i-w}^{i-1} d_{o,j}| - |k - \sum_{o=i-w+1}^{i} d_{o,j}| \leq |d_{i,j} - d_{i-w,j}|$ and $0 < \alpha < 1$, we can obtain,

$$\begin{aligned}&\alpha^{\Delta d} \leq \alpha^{|d_{i,j}-d_{i-w,j}|} \leq \eta \leq \alpha^{-|d_{i,j}-d_{i-w,j}|} \leq \alpha^{-\Delta d}, \\ &e^{-\eta} \leq \eta \leq e^{\eta}. \end{aligned} \tag{13}$$

Thus, $data_{i,j}$ achieves ϵ-differential privacy, and we demonstrate that the differential attack at the cloud server can be prevented.

For each signature $(\sigma_{i,1}, \sigma_{i,2})$, it contains the secret key k_v, such that the fog node and the cloud server cannot recover the individual sensory data report by repetitively trying possible values. Meanwhile, the proposed scheme still supports the sliding window aggregation verification by combining the involved signatures $((\sigma_{i-w+1,1}, \sigma_{i-w+1,2}), \ldots, (\sigma_{i,1}, \sigma_{i,2}))$ and the added noise $(\hat{\sigma}_{i,1}, \hat{\sigma}_{i,2})$. Therefore, the security goal of privacy preservation can be achieved in the proposed scheme.

VI. Performance Evaluation

In this section, we evaluate the performance of the proposed scheme, in terms of complexity and utility. Meanwhile, we compare the proposed scheme with a traditional scheme which adds noise towards each individual data report similar to the process shown in [23], which is briefly described in the following:

- Let $d_{i,j}$ denote the sensory data of dimension j collected during the i-th time slot. Vehicle v selects a random noise $\tilde{r}_{i,j}$ from the truncated α-geometric mechanism $Geom(exp(-\frac{\epsilon}{\Delta d_j}))$ within value range $\{d_j^{min}, d_j^{min} + 1, \ldots, d_j^{max}\}$, and the noise added data is denoted as $\tilde{e}_{i,j} = (d_{i,j} + \tilde{r}_{i,j} + k_{i,j}) \bmod \beta$. Then it structures $\tilde{e}_{i,j}$ and derives the value $\widetilde{data}_i = \sum_{j=1}^{u} \alpha^j \cdot \tilde{e}_{i,j}$.
- To recover the sliding window aggregation result, the value $k'_{i,j}$ is also structured, namely, $\tilde{k}_i = \sum_{j=1}^{u} \alpha^j \cdot k'_{i,j}$. Then the ciphertexts of $(data_i, \widetilde{data}_i, \tilde{k}_i)$ can be derived as

$$\begin{cases} \tilde{c}_{i,1} = g^{data_i} \cdot \tilde{s}_{i,1}^n \mod n^2, \\ \tilde{c}_{i,2} = g^{\widetilde{data}_i} \cdot \tilde{s}_{i,2}^n \mod n^2, \\ \tilde{c}_{i,3} = g^{\tilde{k}_i} \cdot \tilde{s}_{i,3}^n \mod n^2, \end{cases} \tag{14}$$

  where $(\tilde{s}_{i,1}, \tilde{s}_{i,2}, \tilde{s}_{i,3}) \in \mathbb{Z}_n^*$ are three randomly selected numbers. Meanwhile, ciphertext $\tilde{c}_{i,1}$ is designed for the individual sensory data recovery when necessary.
- Vehicle v also selects three random numbers $(\tilde{t}_{i,1}, \tilde{t}_{i,2}, \tilde{t}_{i,3}) \in \mathbb{Z}_q^*$, and computes the signature pairs:

$$\begin{cases} \widetilde{\sigma}_{i,1} = \tilde{t}_{i,1} \cdot P, \\ \widetilde{\sigma}_{i,2} = \tilde{t}_{i,1} \cdot H_1(id_s) + P_{v,0} \\ \qquad + (\sum_{j=1}^{u} d_{i,j} + H_2(k_v||i||0)) \cdot P_{v,1}, \end{cases} \tag{15}$$

  where $(\widetilde{\sigma}_{i,1}, \widetilde{\sigma}_{i,2})$ corresponds to the signature pair of the collected sensory data, and it is introduced for the scenario when the individual sensory data need to be recovered.

$$\begin{cases} \widetilde{\sigma}_{i,3} = \tilde{t}_{i,2} \cdot P, \\ \widetilde{\sigma}_{i,4} = \tilde{t}_{i,2} \cdot H_1(id_s) + P_{v,0} \\ \qquad + (\sum_{j=1}^{u} d_{i,j} + \tilde{r}_{i,j} + H_2(k_v||i||1)) \cdot P_{v,1}, \end{cases} \tag{16}$$

  where $(\widetilde{\sigma}_{i,3}, \widetilde{\sigma}_{i,4})$ corresponds to the signature pair of the collected sensory data at i-th time slot with noise added.

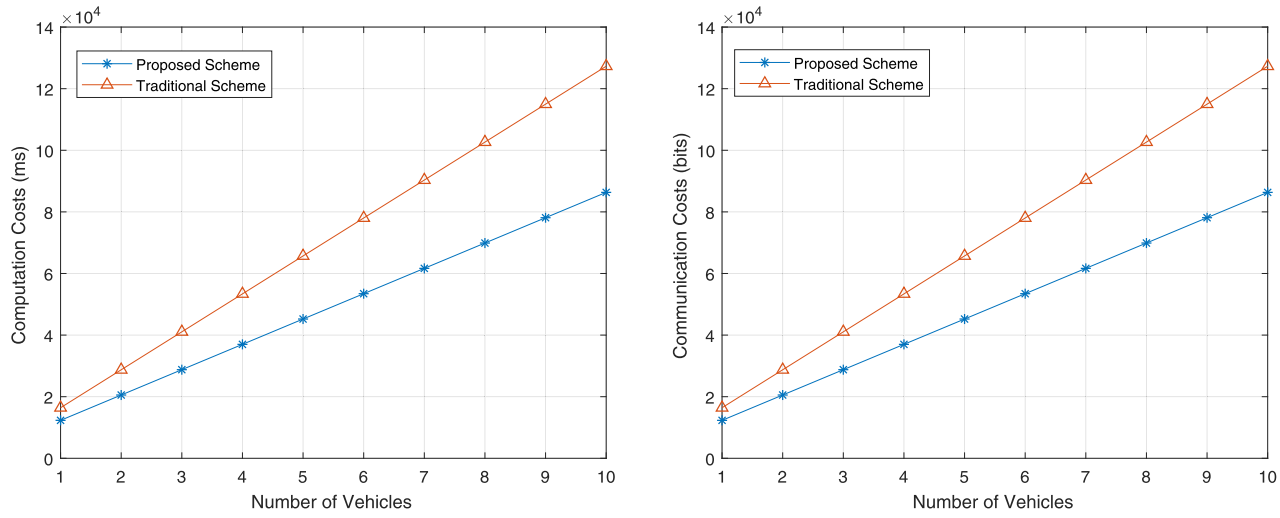
$$\begin{cases} \widetilde{\sigma}_{i,5} = \tilde{t}_{i,3} \cdot P, \\ \widetilde{\sigma}_{i,6} = \tilde{t}_{i,3} \cdot H_1(id_s) + P_{v,0} \\ \qquad + (-\sum_{o=i-w+1}^{i} H_2(k_v||o||1)) \cdot P_{v,1}. \end{cases} \tag{17}$$

  where $(\widetilde{\sigma}_{i,5}, \widetilde{\sigma}_{i,6})$ is utilized for the aggregated sliding window result verification.
- For the fog node and cloud server, they perform the same steps defined in Section IV-C for secure data aggregation, data recovery and authentication processes, respectively.

A. Complexity Analysis

As the computational complexity of the server needs to be involved in the maintenance of all the vehicles, the computation complexity needs to be deliberatively considered. Meanwhile, since the proposed scheme may involve the wireless transmission in the high-mobility scenario, the communication costs also need to be analyzed, and they are reflected through the communication overheads.

1) Computation Cost: When a vehicle id_v generates an encrypted sensory data report $Report_i = id_v||c_i||\hat{c}_i||\sigma_{i,1}||\sigma_{i,2}||\hat{\sigma}_{i,1}||\hat{\sigma}_{i,2}||i$ of time slot i, it requires 4 exponentiation operations in $\mathbb{Z}_{n^2}$ to generate the ciphertext pair $(c_i, \hat{c}_i)$, and 6 multiplication operations in $\mathbb{G}$ to generate the signature pairs $(\sigma_{i,1}, \sigma_{i,2})$ and $(\hat{\sigma}_{i,1}, \hat{\sigma}_{i,2})$. During the data report aggregation phase, as the cost of a multiplication operation in $\mathbb{Z}_{n^2}$ is considered negligible in comparison to the cost of an exponentiation operation, the introduced computation cost for ciphertext aggregation is negligible.



(a) Comparison of Computational Cost

(b) Comparison of Communication Cost

Fig. 3. Comparison of the proposed and the traditional schemes in terms of complexity.

Meanwhile, the cost of an addition operation in $\mathbb{G}$ is also considered negligible in comparison to the cost of a multiplication operation in $\mathbb{G}$. Thus, the introduced computational cost for signature aggregation is also considered negligible. To recover the aggregated sliding window, the cloud server performs 1 exponentiation operation in $\mathbb{Z}_{n^2}$ for decryption, and consumes 4 pairing operations in $\mathbb{G}$ for signature verification.

For the traditional scheme in comparison, it takes 6 exponentiation operations to generate the ciphertexts $(\tilde{c}_{i,1}, \tilde{c}_{i,2}, \tilde{c}_{i,3})$, and 9 multiplication operations in $\mathbb{G}$ to generate signature pairs $(\widetilde{\sigma}_{i,1}, \widetilde{\sigma}_{i,2})$, $(\widetilde{\sigma}_{i,3}, \widetilde{\sigma}_{i,4})$ and $(\widetilde{\sigma}_{i,5}, \widetilde{\sigma}_{i,6})$. Meanwhile, it takes the same steps for the data aggregation and recovery processes, as shown in the above-mentioned analysis.

We denote the computation cost of an exponentiation operation in $\mathbb{Z}_{n^2}$ as c_e, a multiplication operation in $\mathbb{G}$ as c_m, and a pairing operation in $\mathbb{G}$ as c_p. Therefore, the total involved computational cost of the proposed scheme is $(4 * c_e + 6 * c_m) * m + (c_e + 4 * c_b)$, and the computational cost of the traditional scheme is $(6 * c_e + 9 * c_m) * m + (c_e + 4 * c_b)$. We conduct experiments with jPBC [24] and Paillier [25] Libraries on a desktop with a dual core 3.2-GHz processor and an 8-GB installed RAM. Through experiment, we identify that a single exponentiation operation in $\mathbb{Z}_{n^2}$ costs $c_e = 3.7$ ms, a single multiplication operation in $\mathbb{G}$ takes $c_m = 7.95$ ms, and a pairing operation in $\mathbb{G}$ costs $c_p = 4.59$ ms. Fig. 3(a) shows the involved computation cost with respective to the increase of the number of vehicles, when it ranges between 1 to 10. Simulation result shows that our proposed scheme reduces the computation costs in comparison with a traditional scheme.

2) Communication Cost: The proposed scheme has two communication phases: vehicle-to-fog communication and fog-to-cloud communication. For the former phase, the involved communication overhead is $(2048 * 2 + 160 * 4 + 32) * m$ bits, if we set $|id_v| + |i|$ to be 32 bits, and the fog-to-cloud server overhead is $2048 + 160 * 2$ bits. While the vehicle-to-fog communication overhead is $(2048 * 3 + 160 * 6 + 32) * m$ bits, and the fog-to-cloud overhead is $2048 + 160 * 2$ bits. Fig. 3(b) also shows the involved communication overheads with respective to the increase of the number of vehicles, when it ranges between 1 to 10. Simulation result shows that our proposed scheme reduces the communication costs in comparison to the traditional scheme.

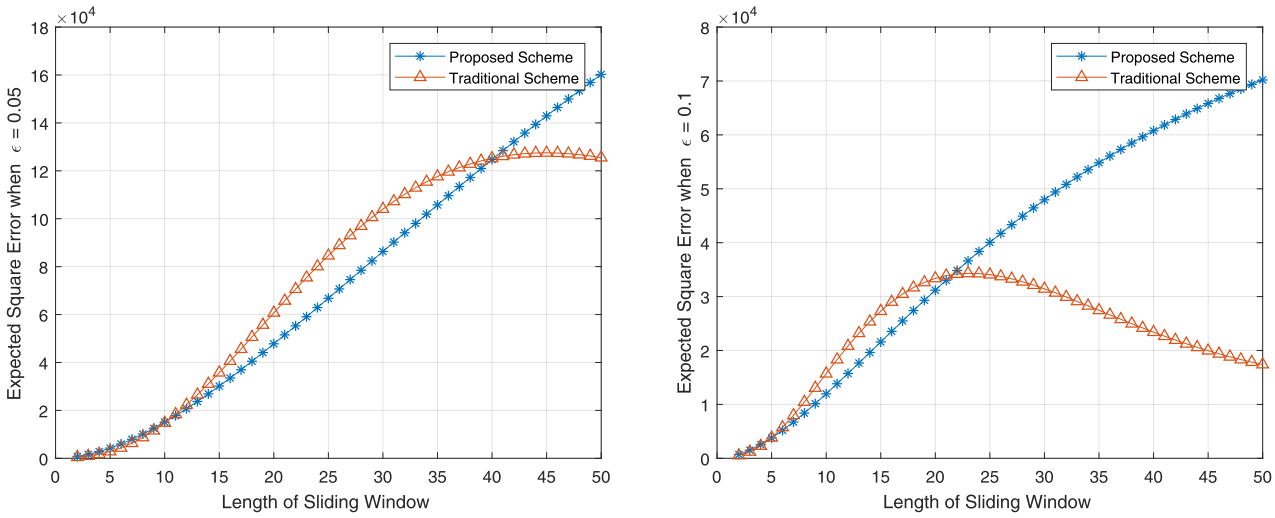
B. Utility Analysis

Since the proposed scheme introduces the noise extracted from the truncated α-geometric mechanism into the aggregation result, we exploit the expected squared error to measure the difference between the noise-added and actual data. In order to investigate the influence of the noise mechanism, we evaluate the proposed scheme and compare it with a traditional scheme under different security level ϵ and the sliding window length.

Specifically, we take one sensory data dimension as an example, the sum of one sensory data dimension is denoted as $f(d) = \sum_{i=1}^{w} d_i$, whose value resides within $\{w \cdot d_{min}, \ldots, w \cdot d_{max}\}$ and the added noise is represented as r_0. Then the expected squared error of the proposed scheme ESE_1 can be denoted as

$$
\begin{aligned}
ESE_1 = E(r_0^2) &= \frac{\sum_{f(d)=w\cdot d_{min}}^{w\cdot d_{max}} E(f(d)^2)}{w \cdot d_{max} - w \cdot d_{min} + 1} \\
\text{where } E(f(d)^2) &= \frac{\alpha^{|w\cdot d_{min} - f(d)|}}{1+\alpha} \cdot (w \cdot d_{min} - f(d))^2 \\
&\quad + \sum_{\delta = w\cdot d_{min} - f(d)+1}^{w\cdot d_{max} - f(d) - 1} \frac{1-\alpha}{1+\alpha} \alpha^{|\delta|} \cdot \delta^2 \\
&\quad + \frac{\alpha^{|w\cdot d_{max} - f(d)|}}{1+\alpha} \cdot (w \cdot d_{max} - f(d))^2.
\end{aligned} \tag{18}
$$

We compare the proposed schemed with a traditional scheme, in which the sensory data value collected during



(a) Expected squared error when $\epsilon = 0.05$

(b) Expected squared error when $\epsilon = 0.1$

Fig. 4. Expected squared error comparison of the proposed and traditional schemes.

time slot i is $g(d) = d_i$, and each vehicle selects a random noise r_i from the truncated α-geometric noise within the range $\{d_{min}, \ldots, d_{max}\}$. Since the proposed scheme considers the sliding window aggregation and any two adjacent aggregation results only differ in one sensory data item, each added geometric random noise needs to follow the ϵ-differential privacy. Thus, the expected squared error of the traditional scheme ESE_2 is

$$ESE_2 = E((\sum_{i=1}^{w} e_i)^2) = w \cdot E(r_i^2) + (w^2 - w) \cdot (E(r_i))^2, \tag{19}$$

where e_i denotes the error introduced by window i, and $E(r_i)$ is derived as

$$\begin{aligned} E(r_i) &= \frac{\sum_{g(d)=n_1}^{n_2} E(g(d))}{d_{max} - d_{min} + 1} \\ &= \sum_{g(d)=n_1}^{n_2} \frac{1}{d_{max} - d_{min} + 1} \left(\frac{\alpha^{|\hat{d}_{min}|}}{1+\alpha} \cdot \hat{d}_{min} \right. \\ &\quad \left. + \sum_{\delta=\hat{d}_{min}+1}^{\hat{d}_{max}-1} \frac{1-\alpha}{1+\alpha} \alpha^{|\delta|} \cdot \delta + \frac{\alpha^{\hat{d}_{max}}}{1+\alpha} \cdot \hat{d}_{max}\right), \end{aligned} \tag{20}$$

where $\hat{d}_{min} = d_{min} - g(d)$ and $\hat{d}_{max} = d_{max} - g(d)$. Then $E(r_i^2)$ can be derived as

$$\begin{aligned} E(r_i^2) &= \frac{\sum_{g(d)=d_{min}}^{d_{max}} E(g(d)^2)}{d_{max} - d_{min} + 1} \\ &= \sum_{g(d)=d_{min}}^{d_{max}} \frac{1}{d_{max} - d_{min} + 1} \left(\frac{\alpha^{|\hat{d}_{min}|}}{1+\alpha} \cdot \hat{d}_{min}^2 \right. \\ &\quad \left. + \sum_{\delta=\hat{d}_{min}+1}^{\hat{d}_{max}-1} \frac{1-\alpha}{1+\alpha} \alpha^{|\delta|} \cdot \delta^2 + \frac{\alpha^{\hat{d}_{max}}}{1+\alpha} \cdot \hat{d}_{max}^2\right). \end{aligned} \tag{21}$$

Fig. 4(a) and Fig. 4(b) compare the expected squared error of proposed scheme and the traditional scheme, with respective to the increase of sliding window length (whose value ranges between 2 to 50), when the privacy level ϵ is set to be 0.05 and 0.1 respectively. Note that the proposed scheme also supports the aggregated sensory data sampling, and we should evaluate the expected squared error under different sliding window length. As shown in Fig. 4(a), the expected squared error of the proposed scheme outperforms that of the traditional scheme, when ϵ is set to be 0.05 and the sliding window length is set between 11 and 40. While as shown in Fig. 4(b), the expected squared error of the proposed scheme outperforms that of the traditional scheme, when ϵ is set to be 0.1 and the sliding window length is set between 5 and 21.

Based on the above complexity analysis in Subsection VI-A and utility analysis in Subsection VI-B, in comparison with the traditional scheme, the proposed scheme reduces the introduced computation and communication costs. While the main limitation of the proposed scheme is the higher expected squared error under certain circumstances, in comparison with the traditional scheme.

VII. Related Work

In this section, we briefly review some works, which are closely related to our work.

A. Privacy-Preserving Data Aggregation

In [11], a privacy-preserving multi-dimensional data aggregation scheme for smart grid was proposed, which utilized the homomorphic Paillier cryptosystem. The proposed scheme enabled the derivation of the aggregated sensory data result of multiple residential user reports, while protecting each individual data report. In [26], an efficient privacy-preserving electricity demand aggregation and response scheme was proposed, by exploiting the homomorphic encryption and

adaptive key evolution techniques. In addition to achieving the aggregation of demands generated by multiple users (such that each individual demand is protected), the proposed scheme also realized the forward session key secrecy and secret key evolutions. In [12], a privacy-preserving location-based data collection and matching scheme was proposed in vehicular ad hoc network, in which the individual vehicular sensory data is denoted in the level of predefined location grid. With the proposed scheme, the individual vehicular sensory data matching result could be obtained, and the individual sensory data report can be protected through data aggregation. A privacy-preserving data collection and querying scheme was proposed in [13], which supports vehicular sensory data collection and moving trajectory representation at the network edge. Through vehicular sensory data aggregation at the network edge, the individual sensory data report could be protected.

The above schemes protect the content of each individual piece of sensory data by aggregating the reports generated by multiple users. However, they are not adaptive to the situation, in which the behavior of each individual user needs to be captured.

B. Differential Privacy in Data Aggregation

To protect the content of each individual piece of sensory data, differential privacy technique was widely utilized for the statistical disclosure control, i.e., assuring each individual user that his/her privacy would not be compromised under certain circumstances.

In [17], a lightweight privacy-preserving data aggregation scheme for fog computing-enhanced IoT was proposed with the homomorphic encryption technique. To achieve the privacy preservation of one IoT device in terms of failure, the proposed scheme adds noise chosen from geometric distribution, such that differential privacy can be achieved in the proposed scheme. In [27], an efficient and privacy-preserving aggregation model with aggregator obliviousness was proposed in smart grid. To achieve differential privacy, it exploits the Gaussian mechanism, and takes both privacy loss mitigation and utility maintenance into consideration. By utilizing the Shamir secret sharing and homomorphic encryption techniques, a differentially private aggregation scheme was proposed for star networks in [18]. This scheme also took collusion into consideration, and guaranteed the correctness of the noise included in the final aggregation result. However, the above schemes are still based on the data aggregation generated by multiple users, and do not take data aggregation in the temporal domain into consideration.

A differential privacy scheme under continual observation was investigated in [28], which was realized through maintaining an accumulative privacy counter. In [20], the authors investigated the scenario when releasing the sliding window data stream aggregation results with differential privacy in the data stream management server. However, the above schemes do not take the distributive data collection architecture into consideration, and they cannot be applied for the decentralized vehicular sensory data collection scenario which makes a brief description of the vehicle's current status. Therefore, a privacy-preserving vehicular sensory sliding window data aggregation scheme with differential privacy is needed.

VIII. Conclusion

In this article, we have proposed a privacy-preserving sliding window based vehicular sensory data collection scheme for predictive maintenance in vehicular fog, which exploits the homomorphic Paillier cryptosystem and the ϵ-differential privacy for the protection of each individual vehicular sensory report. Meanwhile, the proposed scheme achieves the sliding window based aggregation and the correctness verification of the derived aggregation result. Security analysis has demonstrated that our proposed scheme can achieve the predefined security goals, i.e., confidentiality, authentication and privacy preservation. Simulation results have shown that the proposed scheme can reduce the involved communication and computation overheads in comparison to a traditional scheme. For the future work, we will study how to mitigate the performance degradation introduced by the added noise, and further improves the system utility and query accuracy.

References

[1] *Predictive Maintenance for Automobiles*. Accessed: Mar. 10, 2020. [Online]. Available: https://resources.pcb.cadence.com/blog/predictive-maintenance-for-automobiles-2

[2] K. Kaur, S. Garg, G. Kaddoum, E. Bou-Harb, and K.-K.-R. Choo, "A big data-enabled consolidated framework for energy efficient software defined data centers in IoT setups," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2687–2697, Apr. 2020.

[3] S. Grag, K. Kaur, G. Kaddoum, and K.-K. R. Choo, "Toward secure and provable authentication for Internet of Things: Realizing industry 4.0," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4598–4606, May 2020.

[4] Y. Xu, F. Yin, W. Xu, J. Lin, and S. Cui, "Wireless traffic prediction with scalable Gaussian process: Framework, algorithms, and verification," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1291–1306, Jun. 2019.

[5] A. Xie, F. Yin, Y. Xu, B. Ai, T. Chen, and S. Cui, "Distributed Gaussian processes hyperparameter optimization for big data using proximal ADMM," *IEEE Signal Process. Lett.*, vol. 26, no. 8, pp. 1197–1201, Aug. 2019.

[6] K. Kaur, S. Garg, G. Kaddoum, N. Kumar, and F. Gagnon, "SDN-based Internet of autonomous vehicles: An energy-efficient approach for controller placement," *IEEE Wireless Commun.*, vol. 26, no. 6, pp. 72–79, Dec. 2019.

[7] B. Marr, "Big data at VOLVO: Predictive, machine-learning-enabled analytics across petabyte-scale datasets," Bernard Marr Co., Tech. Rep.

[8] A. Chaudhuri, "Predictive maintenance for industrial iot of vehicle fleets using hierarchical modified fuzzy support vector machine," 2018, *arXiv:1806.09612*. [Online]. Available: https://arxiv.org/abs/1806.09612

[9] L. Da Xu, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[10] M. Enev, A. Takakuwa, K. Koscher, and T. Kohno, "Automobile driver fingerprinting," *Proc. Privacy Enhancing Technol.*, vol. 2016, no. 1, pp. 34–50, Jan. 2016.

[11] R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "EPPA: An efficient and privacy-preserving aggregation scheme for secure smart grid communications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 9, pp. 1621–1631, Sep. 2012.

[12] Q. Kong, R. Lu, M. Ma, and H. Bao, "Achieve location privacy-preserving range query in vehicular sensing," *Sensors*, vol. 17, no. 8, p. 1829, Aug. 2017.

[13] Q. Kong, R. Lu, M. Ma, and H. Bao, "A privacy-preserving sensory data sharing scheme in Internet of vehicles," *Future Gener. Comput. Syst.*, vol. 92, pp. 644–655, Mar. 2019.

[14] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. Theory Cryptogr., 3rd Theory Cryptogr. Conf. (TCC)*, New York, NY, USA, Mar. 2006, pp. 265–284.

[15] C. Xu, J. Ren, D. Zhang, and Y. Zhang, "Distilling at the edge: A local differential privacy obfuscation framework for IoT data analytics," *IEEE Commun. Mag.*, vol. 56, no. 8, pp. 20–25, Aug. 2018.
[16] M. Du, K. Wang, X. Liu, S. Guo, and Y. Zhang, "A differential privacy-based query model for sustainable fog data centers," *IEEE Trans. Sustain. Comput.*, vol. 4, no. 2, pp. 145–155, Apr. 2019.
[17] R. Lu, K. Heung, A. H. Lashkari, and A. A. Ghorbani, "A lightweight privacy-preserving data aggregation scheme for fog computing-enhanced IoT," *IEEE Access*, vol. 5, pp. 3302–3312, 2017.
[18] V. Bindschaedler, S. Rane, A. E. Brito, V. Rao, and E. Uzun, "Achieving differential privacy in secure multiparty data aggregation protocols on star networks," in *Proc. 7th ACM Conf. Data Appl. Secur. Privacy*, Scottsdale, AZ, USA, Mar. 2017, pp. 115–125.
[19] F. Kargl, A. Friedman, and R. Boreli, "Differential privacy in intelligent transportation systems," in *Proc. 6th ACM Conf. Secur. Privacy Wireless Mobile Netw. (WiSec)*, Budapest, Hungary, 2013, pp. 107–112.
[20] J. Cao, Q. Xiao, G. Ghinita, N. Li, E. Bertino, and K.-L. Tan, "Efficient and accurate strategies for differentially-private sliding window queries," in *Proc. 16th Int. Conf. Extending Database Technol. (EDBT)*, Genoa, Italy, 2013, pp. 191–202.
[21] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in *Proc. Int. Conf. Theory Appl. Cryptograph. Techn. (EUROCRYPT)*, Prague, Czech Republic, May 1999, pp. 223–238.
[22] C. Gentry and Z. Ramzan, "Identity-based aggregate signatures," in *Proc. Int. Workshop Public Key Cryptogr.*, New York, NY, USA, Apr. 2006, pp. 257–273.
[23] H. Bao and R. Lu, "A new differentially private data aggregation with fault tolerance for smart grid communications," *IEEE Internet Things J.*, vol. 2, no. 3, pp. 248–258, Jun. 2015.
[24] A. De Caro and V. Iovino, "JPBC: Java pairing based cryptography," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2011, pp. 850–855. [Online]. Available: http://gas.dia.unisa.it/projects/jpbc/
[25] *Paillier's Homomorphic Cryptosystem (Java Implementation)*. Accessed: Mar. 10, 2020. [Online]. Available: https://www.csee.umbc.edu/~kunliu1/research/Paillier.html
[26] H. Li, X. Lin, H. Yang, X. Liang, R. Lu, and X. Shen, "EPPDR: An efficient privacy-preserving demand response scheme with adaptive key evolution in smart grid," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 8, pp. 2053–2064, Aug. 2014.
[27] L. Lyu, K. Nandakumar, B. Rubinstein, J. Jin, J. Bedo, and M. Palaniswami, "PPFA: Privacy preserving fog-enabled aggregation in smart grid," *IEEE Trans. Ind. Informat.*, vol. 14, no. 8, pp. 3733–3744, Aug. 2018.
[28] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, "Differential privacy under continual observation," in *Proc. 42nd ACM Symp. Theory Comput. (STOC)*, Cambridge, MA, USA, 2010, pp. 715–724.

Qinglei Kong (Student Member, IEEE) received the B.Eng. degree in communication engineering from the Harbin Institute of Technology, Harbin, China, in 2012, the M.Eng. degree in electronic and information engineering from the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China, in 2015, and the Ph.D. degree from the School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore, in 2018. She used to work at the Cyber Security Cluster, Institute for Infocomm Research, Singapore, and Tencent Security, Shenzhen, as a Research Scientist. She is currently working as a Post-Doctoral Researcher with The Chinese University of Hong Kong (CUHK), Shenzhen. Her research interests include applied cryptography, blockchain, VANET, and game theory.

Rongxing Lu (Senior Member, IEEE) received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo, Canada, in 2012. He worked as an Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University (NTU), Singapore, from April 2013 to August 2016. He worked as a Post-Doctoral Fellow with the University of Waterloo from May 2012 to April 2013. He is currently an Associate Professor with the Faculty of Computer Science (FCS), University of New Brunswick (UNB), Canada. He is a Senior Member of the IEEE Communications Society. He was awarded the most prestigious Governor General's Gold Medal, when he received the Ph.D. degree. He won the 8th IEEE Communications Society (ComSoc) Asia–Pacific (AP) Outstanding Young Researcher Award in 2013. He is the Winner of the 2016–2017 Excellence in Teaching Award in FCS, UNB. He serves as the Vice-Chair (conferences) of the IEEE ComSoc CIS-TC.

Feng Yin (Member, IEEE) received the B.Sc. degree from Shanghai Jiao Tong University, Shanghai, China, in 2008, and the M.Sc. and Dr. Ing. degrees from the Technische Universität Darmstadt, Darmstadt, Germany, in 2011 and 2014, respectively. From 2014 to 2016, he worked at Ericsson Research, Linköping, Sweden, mainly working on the European Union FP7 Marie Curie Training Programme on Tracking in Complex Sensor Systems. He has been working with the Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, since June 2016. His research interests include statistical signal processing, machine learning, and sensory data fusion with applications to wireless positioning and tracking. In 2013, he received the Chinese Government Award for outstanding self-financed students abroad and the Marie Curie Scholarship from the European Union in 2014.

Shuguang Cui (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Stanford University, CA, in 2005. He worked as an Assistant, Associate, Full, and Chair Professor in electrical and computer engineering with The University of Arizona, Texas A&M University, University of California at Davis, and City University of Hong Kong at Shenzhen. He has been the Vice Director of the Shenzhen Research Institute of Big Data. His current research interests include data driven large-scale system control and resource management, large dataset analysis, the IoT system design, energy-harvesting-based communication system design, and cognitive network optimization. He was an Elected Member of the IEEE Signal Processing Society SPCOM Technical Committee from 2009 to 2014 and the Elected Chair of the IEEE ComSoc Wireless Technical Committee from 2017 to 2018. He was a member of the IEEE ComSoc Emerging Technology Committee. He is a member of the Steering Committee for the IEEE TRANSACTIONS ON BIG DATA and the Chair of the Steering Committee for the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He was a recipient of the IEEE Signal Processing Society 2012 Best Paper Award. He was selected as the Thomson Reuters Highly Cited Researcher and listed in the World's Most Influential Scientific Minds by ScienceWatch in 2014. He has served as the general co-chair and the TPC co-chair of many IEEE conferences. He has also served as an Area Editor for the *IEEE Signal Processing Magazine* and an Associate Editor for the IEEE TRANSACTIONS ON BIG DATA, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS Series on Green Communications and Networking, and the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was elected as an IEEE ComSoc Distinguished Lecturer in 2014 and an IEEE VT Society Distinguished Lecturer in 2019.