mathematics

MDPI

Article

Enhancing Real-Time Traffic Data Sharing: A Differential Privacy-Based Scheme with Spatial Correlation

Junqing Le, Bowen Xing, Di Zhang * and Dewen Qiao

College of Computer Science, Chongqing University, Chongqing 400044, China; junqingle@cqu.edu.cn (J.L.); bowenxing@stu.cqu.edu.cn (B.X.); dwqiao@cqu.edu.cn (D.Q.)
* Correspondence: dzhsec@gmail.com

Abstract: The real-time sharing of traffic data can offer improved services to users and timely respond to environmental changes. However, this data often involves individuals' sensitive information, raising substantial privacy concerns. It is imperative to find ways to protect the privacy of the shared traffic data while maintaining its ongoing data utility. In this paper, a Differential Privacy-based scheme with Spatial Correlation for Real-time traffic data (named as DP-SCR) is proposed. DP-SCR not only ensures the high data utility of shared traffic data, but also provides strong privacy protection. Specifically, DP-SCR is designed to adhere to w-event ε-differential privacy, ensuring a high level of privacy protection. Subsequently, a novel adaptive allocation based on spatial correlation prediction is proposed to optimize the privacy budget allocation in differential privacy. In addition, a feasible dynamic clustering algorithm is developed to minimize the relative perturbation error, which further improves the quality of shared data. Finally, the analyses demonstrate that DP-SCR provides w-event privacy for the shared data of each section, and the spatial correlation is a more pronounced characteristic of the traffic data than other characteristics. Meanwhile, experiments conducted on real-world data show that the MAR and MER of the predicted data in DP-SCR are smaller than those in other baseline DP-based schemes. It indicates that the DP-SCR scheme proposed in this paper can provide more accurate shared data.

Keywords: traffic data sharing; privacy protection; differential privacy; adaptive allocation; spatial correlation

MSC: 68P27

Citation: Le, J.; Xing, B.; Zhang, D.; Qiao, D. Enhancing Real-Time Traffic Data Sharing: A Differential Privacy-Based Scheme with Spatial Correlation. *Mathematics* **2024**, *12*, 1722. https://doi.org/10.3390/math12111722

Academic Editor: Faheim Sufi

Received: 5 May 2024
Revised: 27 May 2024
Accepted: 29 May 2024
Published: 31 May 2024

1. Introduction

As science and technology advance, various sensors collect traffic flows (i.e., a kind of traffic statistic) accurately and in real-time [1–6]. Real-time traffic data records the time sequence information of the road and can describe the traffic status in more detail. The real-time traffic data can be shared with other companies and organizations and are subsequently utilized in intelligent transportation systems (ITS), such as traffic light control [7], route planning [8], autonomous driving [9,10], and forecasts of electric vehicle energy consumption [11], ensuring these applications can provide more personalized services and timely respond to environmental changes. However, these traffic statistical data often contain individual sensitive information [12], e.g., location information and vehicle status, which will lead to considerable threats to individual privacy. For example, according to the uniqueness of the individuals' mobility trace [13], an adversary can link back to the individuals' ID information through some outside information when the trace information is published, and then the adversary may match the ID information with sensitive information to acquire individuals' privacy.

To solve the issues of privacy leakage in data sharing, an insightful privacy protection model with strong theoretical support, called ε-differential privacy, has been proposed in [14]. It ensures that the outcomes of any analyses on neighboring datasets

(i.e., two datasets that have only one data difference) are difficult to distinguish. Based on differential privacy, a lot of varietal schemes have been proposed for privacy protection [15–22]. However, most of them focus on either the user-level privacy on finite streams or the event-level privacy on infinite streams. However, applying these methods directly to protect real-time data often leads to inadequate protection or a notable reduction in data utility.

In view of this, Kellaris et al. [23] have proposed a novel model of differential privacy named w-event ε-differential privacy (w-event privacy for short). The w-event privacy model fills the gap between the event-level and the user-level privacy, which can protect all events that happen at any successive w timestamps without sacrificing too much data utility. Using w-event privacy to protect the real-time data is a favorable option. The authors in [23] have designed two schemes based on w-event privacy, called budget distribution (BD) and budget absorption (BA), to protect any event sequence occurring at any successive w timestamps (i.e., a sliding window of size w).

Currently, numerous improved privacy-preserving schemes based on w-event privacy for real-time data sharing have been proposed [24–29]. To improve the accuracy of the shared traffic flows, Wang et al. [24] have proposed two schemes for privacy protection, i.e., RescueDP and E-RescueDP, which take into account data dynamics and can adaptively allocate privacy budgets for each section through proportional-integral-derivative control (PID control) or a recurrent neural network (RNN). Huo et al. [25] have proposed an adaptive w-event privacy for fog computing, which optimizes the prediction of E-RescueDP by using a long short-term memory. In contrast to the centralized differential privacy mentioned earlier, some w-event privacy schemes for real-time data release, based on local differential privacy (abbreviated as LCD), have been developed without the need for establishing a trusted server, as discussed in [27–29].

In this paper, we focus on sharing real-time traffic flows that are used to serve the intelligent transportation system continually. However, if the real-time traffic flows of each road section are shared with the public directly, it can cause serious privacy issues, such as the disclosure of whereabouts. To ensure the shared traffic flows are protected by strong privacy, enabling the sharing of traffic flows that adhere to w-event privacy is essential.

1.1. Motivation

Nevertheless, the prior schemes that focus on real-time data sharing under the protection of w-event privacy have shown limitations in data utility, specifically in terms of the quality of the shared data. Data utility is a vital metric for assessing the quality of the shared data.

First, privacy protection for raw data significantly reduces data utility. In BD and BA schemes [23], they allocate an equivalent privacy budget for the traffic flows of each section. The LCD-based work in [29] divides privacy budgeting to different processing steps to satisfy more limited privacy guarantees. However, the above schemes tend to result in allocating low privacy budgets to traffic flows, which in turn leads to excessive noise introduced into the shared traffic flows. An reasonable allocation of privacy budget is a promising way to solve the above issues. In RescueDP and E-RescueDP in [24], an adaptive allocation is proposed, where the current raw traffic flows are replaced by the predicted traffic flows without consuming privacy budget. In any case, the more accurate the predicted traffic flows are, the more accurate the shared traffic flows also are. However, the calculation of predicted data in [24] is based on the temporal correlation between data, which may not be the best way to predict traffic flows.

Second, the difference in the privacy budget allocated to each section can introduce large relative perturbation error. The sections with small traffic flow produce a large relative perturbation error in the w-event privacy schemes, where the perturbation error is introduced by Laplace noise. To reduce the perturbation error, the mechanism for dynamic grouping in [24] partitions the sections with small traffic flows into different groups, which is based on the similarity of traffic flows. Furthermore, to satisfy w-event privacy, it uses

the smallest privacy budget of the section as the privacy budget of all sections in the group. However, if the sections with a large different privacy budget are partitioned into the same group, it will cause a large perturbation error in all sections of the group. This will lead to a reduction in the accuracy of the shared traffic flows.

1.2. Contributions

Motivated by the above discussions, a scheme named DP-SCR is proposed in this paper to enhance the real-time traffic data sharing. In DP-SCR, we design an adaptive allocation of a privacy budget by using spatial correlation. Then, a novel dynamic clustering method based on k-means algorithm is developed, which takes the traffic flows and the difference in privacy budget into account. Finally, the proposed DP-SCR is proved to satisfy w-event privacy, providing a high level of privacy protection. This means that even if the attacker has background information about the user, they cannot obtain any additional information from the shared data.

Compared with the existing schemes that also satisfy w-event privacy, DP-SCR has the following three contributions:

- In DP-SCR, we prove that the spatial characteristic of traffic flows provides a more remarkable correlation than other characteristics of traffic flows. Then, the designed spatial correlation prediction in DP-SCR is used to adaptively allocate the privacy budget for traffic flows. It significantly improves the accuracy of the shared traffic flows;
- We design a novel dynamic clustering algorithm to aggregate the sections with similar traffic flow and privacy budget. It further improves the accuracy of the shared traffic flows by reducing the relative perturbation error caused by the small traffic flows.
- The experimental results with real-world traffic datasets demonstrate that DP-SCR outperforms baseline w-event privacy schemes in terms of data utility for real-time data release. Also, these experiments validate that DP-SCR is robust to the changes of ϵ and w.

1.3. Organization

The rest of this paper is organized as follows. In Section 2, some preliminary knowledge of the proposed scheme is described. Then, the main problems of the sharing of real-time traffic flows are stated in Section 3. The construction of DP-SCR is established in Section 4, consisting of the adaptive allocation of privacy budgets, dynamic clustering, approximation and perturbation. In Section 5, we analyze the related performance of DP-SCR. The experiments are conducted to verify the high data utility of DP-SCR in Section 6. Finally, the conclusions and future work of this paper can be derived in Section 7.

2. Preliminaries

In this section, we review some basic preliminaries that are necessary for the rest of this paper, mainly including differential privacy, w-event privacy and the characteristics of traffic flows. Some mathematical notations are summarized in Table 1.

Table 1. The mathematical notations.

Notations	Semantic Meanings
M, Q, SC	Laplace mechanism, query function and prediction function, respectively.
S, S_t	An infinite stream and the stream prefix of S at timestamp t, respectively.
D_t	The raw traffic data at timestamp t.
F_i, $F^{i,n}$	The raw traffic flows at timestamp i; the raw traffic flows of section i at successive n timestamps.
f_k^i,$f_k^{x,h}$,$\hat{f}_k^i$	The raw traffic flow of section i at timestamp k; f_k^i at h-th cluster; the predicted traffic flow of section i at timestamp k.

Table 1. *Cont.*

Notations	Semantic Meanings
R_i,R^i,$R^{i,n}$	The sanitized traffic flows at timestamp i, the sanitized traffic flows of section i at all timestamps, and the sanitized traffic flows of section i at successive n timestamps.
r_k^i, $\bar{r}_k^i(out)$	The sanitized traffic flows and the traffic outflow of section i at timestamp k, respectively.
$rn_{i,j}$, RN	The transition probability of that the traffic flows of section i enters to section j; the set of $rn_{i,j}$.
$tp_t^{i,j}$, tp_t^i	The traffic parameter between section i and section j at timestamp t; the set of the traffic parameters between section i and its linked sections at timestamp t.
dis, SSE	The dissimilarity between the predicted traffic flows and the last shared traffic flow; the squared error.
V_{max}^i, V_t^i	The maximal speed limit of section i; the average speed of section i at timestamp t.
$T_{samples}$	The sampling period of raw traffic data.
CA_i	The maximum capacity of section i.
θ_i, ρ	The scale factor of corrections.
C_t^i, c_t^i	The i-th cluster at timestamp t and the cluster center of C_t^i, respectively.
CLU_t	The set of clusters at timestamp t.
ε_t^i, $\varepsilon_t^{x,i}$, $\hat{\varepsilon}_i^j$	The privacy budget of section i at timestamp t; the privacy budget included in C_t^i; the privacy budget of C_t^i.
ε_r, ε_{max}	The remaining privacy budget; the maximum privacy budget allowed for sections.
λ_i^j	The perturbation error of section i at timestamp j.

2.1. Differential Privacy

Let $\mathcal{D}$ denote a set of datasets, and let Q be the query function. M represents the Laplace mechanism, and the set R denotes the range of $M(\cdot)$.

Definition 1 (Neighboring datasets [30]). *For two datasets $D \in \mathcal{D}$ and $D' \in \mathcal{D}$, if D' can be obtained from D by removing or adding any single record, D and D' are neighboring.*

Definition 2 (Sensitivity [30]). *Assume $Q : \mathcal{D} \to \Re^d$, then the sensitivity of Q with regard to $\mathcal{D}$ is*

$$\Delta(Q) = \max_{D,D'} \parallel Q(D) - Q(D') \parallel_1,$$

where D and D' represent any pair of neighboring datasets of $\mathcal{D}$.

Definition 3 (Laplace mechanism [30]). *For $Q : \mathcal{D} \to \Re^d$, M adds noise into the results of $Q(D)$, where the noise conforms to the Laplace distribution. Formally, for any dataset $D \in \mathcal{D}$,*

$$M(D) = Q(D) + \langle Lap(\Delta(Q)/\varepsilon) \rangle^d,$$

where ε denotes privacy budget indicating the privacy level of mechanism M.

Definition 4 (Differential privacy [31]). *For any neighboring dataset D and D', and the set R, if*

$$Pr[M(D) \in R] \leq e^{\varepsilon} Pr[M(D') \in R],$$

the mechanism M satisfies ε-differential privacy ($\varepsilon > 0$).

Take the traffic flows of section k at timestamp i as an sample. The queried result of the traffic flows from D_i is represented as $Q_k(D_i) = f_i^k$, and the shared traffic flows after

the processing of differential privacy can be rewritten as $r_i^k = f_i^k + \langle Lap(\Delta(Q_k)/\varepsilon)\rangle$, where D_i is the raw traffic data at timestamp i and Q_k is the query function for section k.

Theorem 1 (Sequential composition [32]). *Assume that M includes a sequence of sub-mechanisms $M_1, M_2, \cdots, M_r$ and each M_i adds an independently random noise. If each mechanism M_i satisfies ε_i-differential privacy, the mechanism M satisfies $(\sum_{i=1}^{r} \varepsilon_i)$-differential privacy.*

According to the above definitions and Theorem 1, it is obvious that the smaller ε or the higher $\Delta(Q)$ is, the larger the noise introduced. The privacy budget ε of M assigned to sub-mechanisms may be different.

2.2. w-Event Privacy

w-event privacy can protect all events that happen at any successive w timestamps. For a distinct description of w-event privacy, the traffic data are denoted as an infinite tuple $S = (D_1, D_2, \cdots)$, where D_t represents the raw traffic data at timestamp t, and $S[i]$ is the i-th element of S. Then a stream prefix of S at timestamp t is denoted as $S_t = (D_1, D_2, \cdots, D_t)$.

Definition 5 (w-neighboring [23]). *Two stream prefixes S_t and S'_t are w-neighboring, where w is a positive integer, if they satisfy the following conditions:*

1. *For each $S_t[i]$, $S'_t[i]$ with $i \in [t]$ and $S_t[i] \neq S'_t[i]$, it holds that $S_t[i]$, $S'_t[i]$ are neighboring;*
2. *For each $S_t[i_1]$, $S_t[i_2]$, $S'_t[i_1]$, $S'_t[i_2]$, when $i_1 < i_2$, $S_t[i_1] \neq S'_t[i_1]$ and $S_t[i_2] \neq S'_t[i_2]$, it holds that $i_2 - i_1 + 1 \leq w$;*

Definition 6 (w-event privacy [23]). *Let $S_t[i] = D_i \in \mathcal{D}$ and one set $R \subset Range(M)$. For all w-neighboring stream prefixes S_t, S'_t and all t, if*

$$Pr[M(S_t) \in R] \leq e^{\varepsilon} Pr[M(S'_t) \in R],$$

the mechanism M satisfies w-event privacy.

Theorem 2. *Let stream prefix S_t denote the input of M, and the output of M is $\{R_1, R_2, \cdots, R_t\} \subset Range(M)$. Suppose that the mechanism M includes t mechanisms $M_1, M_2, \cdots, M_t$, and each $M_i(S_t[i])$ achieves ε_i-differential privacy. Then the mechanism M satisfies w-event privacy, if*

$$\forall i \in [t], \sum_{k=i-w+1}^{i} \varepsilon_k \leq \varepsilon.$$

2.3. Characteristics of Traffic Flows

Based on the analyses in Section 1,the proposed DP-SCR mainly considers the spatial correlation between traffic flows.

Definition 7 (Spatial correlation [33]). *A road network consists of multiple sections, and there exists a spatial correlation between sections. Formally, the algorithm SC denotes the spatial correlation between the traffic flows. The predicted traffic flow $\hat{f}_{t+1}^i$ can be calculated by $\hat{f}_{t+1}^i = SC(tp_t^i)$, where tp_t^i is the traffic parameter of section i at timestamp t. If the linked sections of section i are section j and section k, then $tp_t^i = \{tp_t^{i,j}, tp_t^{i,k})\}$, where $tp_t^{i,k}$ consists of the shared traffic flow $r_t^{i,k}$ and some prior knowledge including the maximal speed limit v_{max}^i, the predefined sampling period, and the road networks $rn_{k,i}$.*

The road networks link with all the sections and show the flow correlation between different sections. As shown in Figure 1, it is a part of the road networks, where each numeric value represents the probability that the traffic flow of one section enters its linked (adjacent) sections.

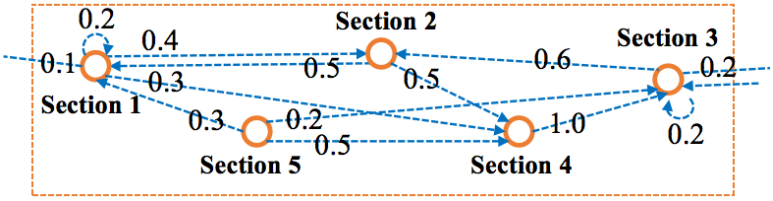

Figure 1. Partial road networks and its transition probabilities.

For example, there are half traffic flows of Section 5 that enter to Section 4, so the probability is 0.5 and is denoted as $rn_{5,4}$ in this paper. Obviously, the flow of traffic on any section will either stay in the same section or enter to another section, so the probability of section i satisfies the following relationship.

$$\sum_{j \in A_i} rn_{i,j} = 1,$$

where A_i is a set includes section i and its linked sections. The flow correlation is represented as $RN = \{rn_{i,j} | i, j \in Sec\}$, where Sec is the set of all sections.

Mathematical notations: The mathematical notations and their semantic meanings used in this paper are summarized in Table 1.

3. Problems Statement

When real-time traffic flows are shared with the public, they may cause serious privacy issues. Therefore, in order to ensure the shared traffic flows with strong privacy protection, each section is required to satisfy w-event privacy. As shown in Figure 2, the traffic data are collected by various sensors and stored in the database. Then, the traffic flows for serving an intelligent transportation system will be processed to satisfy w-event privacy, so that the shared flows can not leak the privacy of users. To be more specific, let F_k be the traffic flows of section k and D_k be raw traffic data at timestamp k. Then, we have $F_k = Q(D_k) = (f_k^1, f_k^2, \cdots, f_k^n)$, where n is the total number of sections at timestamp t, and f_k^i is defined as the traffic flow of section i at timestamp k. In order to ensure the traffic flows are shared securely, the sanitized version of f_k^i, denoted by r_k^i, is used to replace f_k^i. Thus, the sanitized version of infinite time traffic flows at section i is denoted as $R^i = (r_1^i, r_2^i, \cdots, r_k^i, \cdots)$.

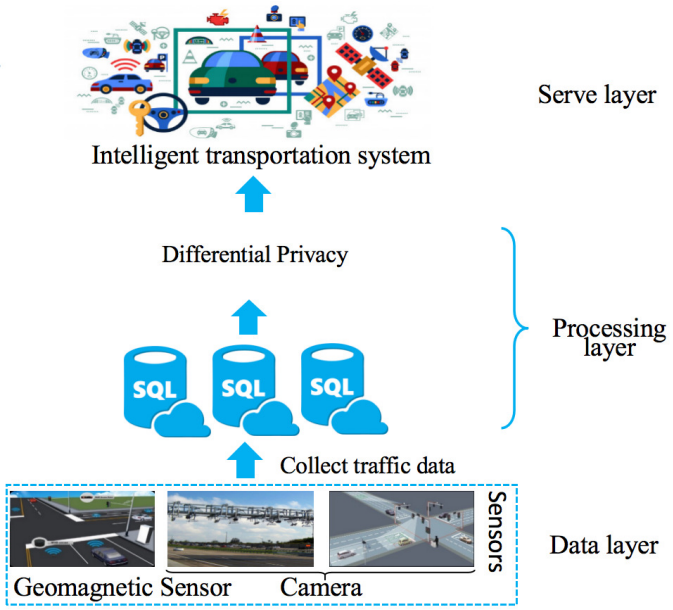

Figure 2. The system model.

In this paper, the problem concerning privacy protection is formally stated as follows.

Given an infinite time series of traffic flows $\mathbb{F} = (F_1, F_2, \cdots, F_k, \cdots)$, denote its sanitized version as $\mathbb{R} = (R_1, R_2, \cdots, R_k, \cdots)$. Then, a scheme is designed to make each infinite time section from $\mathbb{R}$, denoted as $R^i = (r_1^i, r_2^i, \cdots, r_k^i, \cdots)$, which satisfies w-event privacy.

Since data utility is the main criterion for measuring the quality of a scheme, designing a mechanism to improve the data utility of the shared traffic flows is very meaningful. In this paper, the allocation of the privacy budget and the perturbation error will affect the accuracy of the shared traffic flows greatly. Therefore, the problem of the data utility can be described as follows.

(a) How to allocate the privacy budget reasonably. (b) How to reduce the absolute error (MAE) and relative error (MRE) of the shared traffic flows $\mathbb{R}$, where MAE and MRE are the representation of perturbation error.

4. The Design of DP-SCR

In this section, we propose a scheme, named DP-SCR, which satisfies w-event privacy and provides high accuracy of traffic flows. The proposed DP-SCR can achieve the adaptive allocation of privacy budget, where the spatial correlation prediction is used to improve the budget allocation. Additionally, dynamic clustering is proposed to reduce the perturbation error caused by the small traffic flows. Finally, a novel approximation method and the perturbation method are used to deal with the no sampled sections and sampled sections in DP-SCR, respectively. Figure 3 shows the flowchart of DP-SCR, where the sampling of sections is determined by dis and λ_{i+1}. The dis is the dissimilarity between the predicted traffic flow and the last shared traffic flow, and λ_{t+1} is perturbation error.

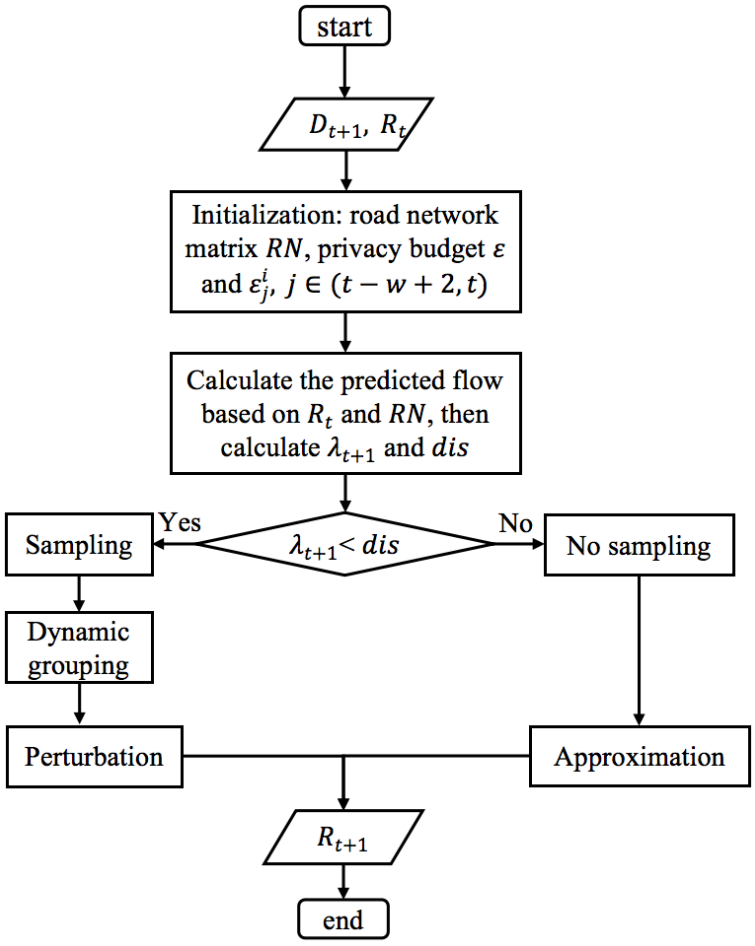

Figure 3. The flowchart of DP-SCR.

Algorithm 1 gives an overall description of the proposed DP-SCR. The main processes of DP-SCR are described in detail as follows.

Algorithm 1: DP-SCR.

Require: raw traffic data D_{t+1}, the shared traffic flows R_t, $tp_t = \{tp_t^i, \cdots, tp_t^n\}$.
Ensure: new shared traffic flows R_{t+1}.
1: Obtain traffic flows $F_{t+1} = Q(D_{t+1})$;
2: **for** each section i at timestamp t **do**
3: Calculate the predicted traffic flow $\hat{f}_{t+1}^i$ according to tp_t^i, and then calculate dis;
4: Calculate the privacy budget for each section at timestamp t;
5: Sampling according to $\hat{f}_{t+1}$.
6: **end for** (*see Section 4.1*)
7: For sampling points, do dynamic clustering for them at timestamp $t+1$, and perturb their traffic flows by adopting Laplace mechanism; (*see Sections 4.2 and Section 4.3*)
8: For non-sampled points, approximate current traffic flows with the corresponding predicted traffic flows; (*see Section 4.3*)
9: Obtain R_{t+1} by combining the results at sampling points and the results at non-sampled points.

4.1. Adaptive Allocation of Privacy Budget

According to Theorems 1 and 2, if the sliding window size w is too large, the privacy budget allocated for the sections at each timestamp is small, which will result in a large magnitude of noise. Sampling is a promising way to reduce noise. Because the non-sampled points do not consume any privacy budget, more privacy budget will be allocated for the sampling points. Towards this end, an adaptive allocation of the privacy budget is proposed in the literature [24]. It adopts the temporal correlation to predict the value at the next timestamp, and the value is used for sampling, where the predicted value determines the quality of the sampling. Also, due to the spatial correlation between traffic flows, it can increase the accuracy of predicted traffic flows and make the sampling more reasonable.

Inspired by the above ideas, a mechanism for the adaptive allocation of the privacy budget based on spatial correlation is proposed in this paper. The mechanism includes three operations, which are described in detail as follows.

4.1.1. Spatial Correlation Prediction

In the phase of spatial correlation prediction, we note that V_{max}^i is the maximal speed limit of section i, where the speed of all vehicles is assumed to be less than or equal to the maximum speed. The predefined T_{Sample} is the sampling period of raw traffic data. According to Equation (9) of the work [34], the speed–flow relationship between V_t^i and r_t^i is $V_t^i = \theta_1 \times V_{max}^i / (1 + (r_t^i / CA_i)^\rho)$, $\rho = \theta_2 + \theta_3 \times (r_t^i / CA_i)^3$, where V_t^i is the average vehicle speed in section i at timestamp t, $CA_i = V_{max}^i \times T_{Sample}$ is the maximum capacity of section i, and θ_i $\{i = 1, 2, 3\}$ and ρ are scale factor corrections. It is obvious that the average vehicle speed is related to the shared traffic flows, the maximal speed limit of the section, and the predefined sampling period. However, the scale factors are hard to set artificially. Inspired by [34], we design a novel spatial correlation algorithm SC to calculate the predicted traffic flows of section i. The goal of the method is to predict the traffic flow of section i at timestamp $t+1$ based on r_t^i and the prior knowledge V_{max}^i, T_{Sample} and $rn_{k,i}$. Specific processes are described in the following three steps:

Step 1: To calculate V_t^i, we train a model M_v to learning the relationship between V_t^i and r_t^i, V_{max}^i and T_{Sample}. Based on the trained model, we can obtain the average vehicle speed of each section at any timestamp by inputing the prior knowledge and the shared traffic flows.

Step 2: Based on the average speed of section i and the sampling period T_{Sample}, the traffic outflow $\bar{r}_t^i(out)$ of section i is calculated by

$$\bar{r}_t^i(out) = V_t^i \times T_{Sample}.$$

Step 3: According to the actual situation, the predicted traffic flow ($\hat{f}_{t+1}^i$) of section i is the difference between the traffic outflow of section i and the traffic inflows of its linked sections. For example, if section j and section k are the linked sections of section i, the predicted traffic flow of section i is represented as the following formula.

$$\hat{f}_{t+1}^i = rn_{j,i} \cdot \bar{r}_t^j(out) + rn_{k,i} \cdot \bar{r}_t^k(out) - (1 - rn_{i,i}) \cdot \bar{r}_t^i(out).$$

4.1.2. Calculation of Privacy Budget

In the calculation of the privacy budget, to satisfy w-event privacy, the total privacy budget of each section at any sliding window should be smaller than ε. Here, assume that all sections at the next timestamp are sampling points. Thus, the privacy budget of all sections at the next timestamp should be calculated.

Without loss of generality, let the current timestamp be t; then, the privacy budget for section i at timestamp $t + 1$ is ε_{t+1}^i. The remaining privacy budget ε_r in the sliding window $[t - w + 2, t]$ is calculated by $\varepsilon_r = \varepsilon - \sum_{j=t-w+2}^{t} \varepsilon_j^i$. Additionally, the sampling interval is $I = (t + 1 - l)$, where l is the last sampling point of section i. Then, a scale factor p, which determines how much privacy budget will be allocated for section i at timestamp $t + 1$, is calculated by

$$p = \min(\varphi \times ln(I + 1), p_{max}),$$

where φ is defined as a scale factor varied in (0, 1], and p_{max} is the maximum portion of privacy budget allocated for each sampling point. In the end, the privacy budget allocated for section i at timestamp $t + 1$ is calculated by

$$\varepsilon_{t+1}^i = \min(p \times \varepsilon_r, \varepsilon_{max}),$$

where ε_{max} is the maximum privacy budget allocated for each sampling point. Two constraints (i.e., p_{max} and ε_{max}) are aimed at striking a good balance between the data utility and privacy protection of traffic flows.

4.1.3. Sampling with the Predicted Traffic Flows

The perturbation error of section i is $\lambda_{t+1}^i = 1/\varepsilon_{t+1}^i$, and the dissimilarity between the predicted traffic flow and the last shared traffic flow is $dis = \hat{f}_{t+1}^i - r_l^i$, where r_l^i is the last shared traffic flow of section i. If $\lambda_{t+1}^i > dis$, the traffic flow at timestamp $t + 1$ is approximated by the predicted traffic flow. Then, the privacy budget of section i is withdrawn, i.e., section i at timestamp $t + 1$ is a non-sampled point, and its privacy budget is zero. Otherwise, section i at timestamp $t + 1$ is a sampling point, and its privacy budget remains unchanged. The mechanism for the adaptive allocation of the privacy budget is formally presented in Algorithm 2.

Algorithm 2: Adaptive allocation of privacy budget for section i at timestamp $t+1$.

Require: privacy budget ε, ε_{max}, p_{max}, r_l^i, RN and the traffic flows of section i and its linked sections at timestamp t.

Ensure: the privacy budget of section i at timestamp $t+1$.

1: Assume that section i at timestamp $t+1$ is sampling point, and calculate privacy budget for it, then obtain $\varepsilon_{t+1}^i = min(p \times \varepsilon_r, \varepsilon_{max})$, where $p = min(\varphi \times ln(I+1), p_{max})$ and $i = (t+1-l)$;
2: According to *Spatial Correlation Prediction*, calculate $\hat{f}_{t+1}^i$ that is the predicted traffic flow of section i at timestamp t +1;
3: Calculate the dissimilarity between the predicted traffic flow and the last sharing $dis = \hat{f}_{t+1}^i - r_l^i$;
4: $\lambda_{t+1}^i = 1/\varepsilon_{t+1}^i$;
5: **if** $dis > \lambda_{t+1}^i$ **then**
6: section i at timestamp $t+1$ is sampling point;
7: return ε_{t+1}^i;
8: **else**
9: section i at timestamp $t+1$ is non-sampled point;
10: return 0;
11: **end if**

4.2. Dynamic Clustering

As shown in the analyses in Section 1, the sections with small traffic flow can cause large relative perturbation error in the w-event privacy schemes. In this section, a dynamic clustering algorithm, i.e., bisecting k-means, is adopted to reduce the perturbation error. Specifically, the sections with similar traffic flow and privacy budget will be aggregated together to resist noise via the dynamic clustering.

First, it is necessary to determine which sections have small traffic flow before clustering. Here, the noise resistance threshold is defined as τ, which reflects whether the traffic flows have sufficient capacity to resist noise. When the traffic flows are smaller than τ, they are classified as small traffic flows. Then, the sections with small traffic flows will be saved in the cluster C_{t+1}^0.

Assume that the number of the sections with small traffic flow is n; then $C_{t+1}^0 = \{y_{1,0}, \cdots, y_{n,0}\}$, $y_{x,0} = (f_{t+1}^{x,0}, \lambda_{t+1}^{x,0})$, and $\lambda_{t+1}^{x,0} = 1/\varepsilon_{t+1}^{x,0}$, $x \in [1,n]$, where $f_{t+1}^{x,0}$ is the traffic flow f_{t+1}^x of cluster C_{t+1}^0, and $\lambda_{t+1}^{x,0}$ is perturbation error. When $\sum_{x \in C_{t+1}^h} f_{t+1}^{x,h} \geq \tau$, the cluster at timestamp $t+1$ can be denoted as $CLU_{t+1} = \{C_{t+1}^0, \cdots, C_{t+1}^k\}$, $(k \leq n)$. Also, the sum of the squared error (SSE) of CLU_{t+1} is $SSE(CLU_{t+1}) = \sum_{h=1}^{k} \sum_{x \in C_{t+1}^h} ||y_{x,h} - c_{t+1}^h||^2$, where c_{t+1}^h is the cluster center of C_{t+1}^h. As is well known, the smaller the SSE is, the more similar the traffic flows and privacy budget of sections are. Thus, the dynamic clustering is aimed at finding the smallest SSE, which is described in Algorithm 3.

After dynamic clustering, suitable privacy budget should be allocated for each cluster in CLU_{t+1}. Without loss of generality, $\sum_{j=0}^{w-1} \varepsilon_{t+j}^i$ is denoted as the total privacy budget for section i at any successive w timestamps. In order to ensure $\sum_{j=0}^{w-1} \varepsilon_{t+j}^i \leq \varepsilon$, the privacy budget allocated for C_{t+1}^i is equal to $\hat{\varepsilon}_{t+1}^i$, and the privacy budget of the sections in C_{t+1}^i is also $\hat{\varepsilon}_{t+1}^i$, where $\hat{\varepsilon}_{t+1}^i = \min_{x \in C_{t+1}^i} (\varepsilon_{t+1}^{x,i})$.

Algorithm 3: Dynamic Clustering Algorithm at timestamp $t+1$.

Require: C^0_{t+1}.
Ensure: CLU_{t+1}.
Initialization: $Clu = C^0_{t+1}$, $\sum_{x \in C^0_{t+1}} f^{x,0}_{t+1} \leq \tau$.
if $\sum_{x \in C^0_{t+1}} f^{x,0}_{t+1} \leq \tau$ **then**
  return Clu;
end if
while 1 **do**
  $k = size(Clu)$; $Clu_2 = \{C^0_{t+1}, \cdots, C^k_{t+1}\}$;
  for $i = 0 : k$ **do**
    do 2-*means* for C^i_{t+1} in Clu, then obtain new C^i_{t+1} and C^{k+1}_{t+1};
    if $\sum_{x \in C^i_{t+1}} f^{x,i}_{t+1} \geq \tau$ and $\sum_{x \in C^{k+1}_{t+1}} f^{x,k+1}_{t+1} \geq \tau$ **then**
      $Clu_1 = \{C^0_{t+1}, \cdots, C^i_{t+1}, \cdots, C^{k+1}_{t+1}\}$; (the clusters except C^i_{t+1} and C^{k+1}_{t+1} are from Clu)
      if $SSE(Clu_1) < SSE(Clu_2)$ **then**
        $Clu_2 = Clu_1$
      end if
    end if
  end for
  if $Clu \neq Clu_2$ **then**
    $Clu = Clu_2$;
  else
    return Clu;
  end if
end while
The function of 2-means is as follows.
Function: $(C^i_{t+1}, C^{k+1}_{t+1})$=2-*means*$(C^i_{t+1})$
1: randomly select two objects from these k objects of C^i_{t+1} as the initial cluster centers of the cluster A and the cluster B;
2: calculate the similarity (Euclidean distance) between each object $y_{x,i}$ and the cluster center; (The smaller the value is, the closer the similarity is)
3: all objects are divided into the cluster with closer similarity;
4: recalculate the cluster centers of the cluster A and the cluster B;
5: repeat step 2–step 4 until each cluster is not changing;
6: $C^i_{t+1} = A$, and $C^{k+1}_{t+1} = B$;
7: return C^i_{t+1} and C^{k+1}_{t+1}.
end Function

4.3. Approximation and Perturbation

To ensure that each section satisfies w-event privacy, the noise that conforms to Laplace distribution is injected into each sampling section. In [24], it uses the last shared value to approximate non-sampled sections. Different from the approximation mechanism in [24], we propose a novel approximation mechanism that takes the predicted value as the value of non-sampled sections. The predicted values are used to approximate the real value in this paper. However, there exists a dissimilarity *dis* between the last shared values and the predicted values, so the predicted values are closer to the real traffic flows of the section than to its last shared value. In any case, the predicted values are calculated based on the shared traffic flow at the previous timestamp. Thus, it also can protect real values and prevent privacy leakage.

In the perturbation mechanism, D_{t+1} is the raw traffic data at timestamp $t+1$, and C^h_{t+1} is a cluster at timestamp $t+1$ consisting of n_h sections. As each vehicle can only appear in at most one section at each timestamp, the sensitivity of Q ($\Delta(Q)$) is 1. Then, the sanitized traffic flow of section i at timestamp $t+1$ can be denoted as

$$M(D^i_{t+1}) = \begin{cases} Q(D^i_{t+1}) + Lap(\Delta(Q)/\varepsilon^i_{t+1}), \ if \ i \notin C^h_{t+1} \\ (Q(D^i_{t+1}) + Lap(\Delta(Q)/\hat{\varepsilon}^i_{t+1}))/n_h, \ otherwise. \end{cases}$$

If section $i \notin C^h_{t+1}$, the sanitized traffic flows are $Q(D^i_{t+1}) + Lap(\Delta(Q)/\varepsilon^i_{t+1})$. Otherwise, the sanitized traffic flows are $(Q(D^i_{t+1}) + Lap(\Delta(Q)/\hat{\varepsilon}^i_{t+1}))/n_h$.

5. Performance Analyses

In this section, we will analyze the privacy protection, the correlation of traffic flows and effects of filtering in DP-SCR.

5.1. Privacy Analyses

In this subsection, the privacy loss and privacy protection are analyzed.

5.1.1. Privacy Loss

Privacy loss is used to metric the privacy information leakage. According to the definition of differential privacy, we have the privacy loss

$$\begin{aligned} &\ln \frac{\Pr(M(D) = r)}{\Pr(M(D') = r)} = \ln \frac{\Pr(Q(D) + \langle Lap(\Delta(Q)/\varepsilon)\rangle^d = r)}{\Pr(Q(D') + \langle Lap(\Delta(Q)/\varepsilon)\rangle^d = r)} \\ &= \ln \frac{\Pr(\langle Lap(\Delta(Q)/\varepsilon)\rangle^d = r - Q(D))}{Pr(\langle Lap(\Delta(Q)/\varepsilon)\rangle^d = r - Q(D'))} \\ &= \ln \frac{exp(-|r - Q(D)|\varepsilon/\Delta(Q))}{exp(-|r - Q(D')|\varepsilon/\Delta(Q))} \\ &= \ln exp(\varepsilon(|r - Q(D)| - |r - Q(D')|)/\Delta(Q)) \\ &\leq \ln exp(\varepsilon|Q(D) - Q(D')|/\Delta(Q)) \\ &\leq \ln exp(\varepsilon) \\ &\leq \varepsilon \end{aligned}$$

where r represents the output after the processing of differential privacy. Therefore, the privacy loss is determined by the allocated privacy budget ε and does not exceed ε.

5.1.2. Privacy Protection

These schemes BA [23], BD [23], E-RescueDP [24] and CLDP [29] to be compared in this paper all satisfy w-event privacy. Here, we will prove whether the DP-SCR proposed in this paper satisfies w-event privacy.

Claim 1. *The proposed DP-SCR satisfies w-event privacy.*

Proof. In DP-SCR, the perturbation phase is the only one accessing raw traffic flow. If $\sum_{j=t-w+1}^{t} \bar{\varepsilon}^i_j \leq \varepsilon$ for each section, Claim 1 holds, where $\bar{\varepsilon}^i_j$ is the privacy budget allocated for section i at timestamp j in perturbation.

For section i, ε^i_t is the allocated privacy budget at timestamp t after the adaptive allocation of the privacy budget. Then, the budget privacy ε^i_t will be changed after dynamic clustering, and the changed privacy budget will be denoted as $\bar{\varepsilon}^i_t$. If section i belongs to C^h_t, $\bar{\varepsilon}^i_t = \hat{\varepsilon}^h_t$, where $\hat{\varepsilon}^h_t = \min_{x \in C^h_t}(\varepsilon^{x,i}_t)$, meaning $\varepsilon^i_t \geq \hat{\varepsilon}^h_t$. Otherwise, $\bar{\varepsilon}^i_t = \varepsilon^i_t$. Thus, $\varepsilon^i_t \geq \bar{\varepsilon}^h_t$ holds.

Moreover, as $\sum_{j=t-w+1}^{t} \varepsilon_j^i \leq \varepsilon$ for section i at any successive w timestamps has been required in the mechanism of the adaptive allocation of the privacy budget, $\sum_{j=t-w+1}^{t} \bar{\varepsilon}_j^i \leq \varepsilon$ is tenable. Finally, according to Theorem 2, the proposed DP-SCR satisfies w-event privacy, so Claim 1 holds. □

5.2. Correlation Analyses

Claim 2. *The spatial correlation between traffic flows is more remarkable than other characteristics of traffic flows in the prediction.*

Proof. Traffic flows have four characteristics, i.e., temporal correlation, spatial correlation, historical correlation and multistate. The authors in [33] indicated that the prediction for traffic flows is only related to the temporal, spatial and historical correlation between traffic flows, and they illustrated that the multistate is useless. Upon further analysis, the prediction for traffic flows also has little effect on the historical correlation between traffic flows due to the following reasons: (1) On-road traffic events, such as accidents and road closures, affect the traffic flows in the transportation system, and these effects cannot be predicted a priori [35]. (2) Off-road events have a major impact on the traffic flows and may not be included in the usual historical traffic flows [35]. (3) The timestamps of sampling are too short to predict the traffic flows at the next timestamp by using historical traffic flows in the sharing of real-time traffic flows. Thus, the prediction for traffic flows is mainly correlated with temporal and spatial characteristics. Also, the authors in [36,37] have also emphasized that most of the mechanisms on the prediction for traffic flows mainly are based on temporal correlation and spatial correlation. However, the spatial characteristics of traffic flows can reflect the correlation between traffic flows more distinctly than the temporal characteristics of traffic flows, which has been illustrated as follows.

The Pearson correlation coefficient is used to calculate the spatial correlation between traffic flows, i.e., $\rho_{X,Y} = COV(X,Y)/\sigma_X\sigma_Y$, where COV is covariance, σ_X and σ_Y are, respectively, the standard deviation of X and Y. In the experiment, the traffic flows of the target section and the traffic flows of its linked sections at 160 successive timestamps are, respectively, served as X and Y. Then, the spatial correlation between traffic flows is 0.7072. Also, the autocorrelation coefficient is adopted to calculate the temporal correlation of the above-mentioned traffic flows, where the range of retardation timestamp is [0, 10], and the sample size is 10,000, which is large enough for the coefficient calculation. Figure 4 shows the temporal correlation of X, where all results are smaller than 0.3. As the larger correlation value means a more remarkable correlation, the spatial correlation between traffic flows is more striking than the temporal correlation between traffic flows.

That is, the prediction based on spatial correlation obtains more accurate results than that based on other characteristics of traffic flows. Thus, Claim 2 holds. □

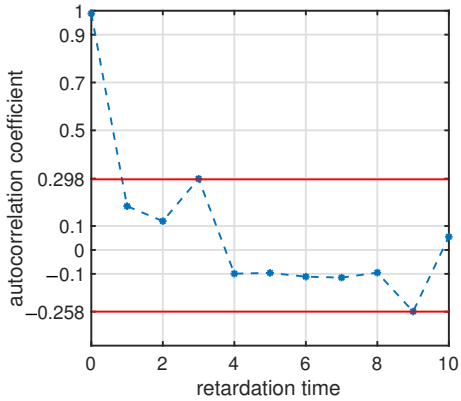

Figure 4. The temporal correlation (autocorrelation) of X.

5.3. Effects of Filtering on DP-SCR

In many differential privacy schemes, the sanitized traffic flows can not be shared directly because the noise caused by perturbation may reduce the accuracy of the shared traffic flows. Thus, the filtering mechanism is used to improve the accuracy of the sanitized traffic flows after perturbation.

In E-RescueDP [24], it uses Kalman Filter to improve the accuracy of the sanitized traffic flows. To compare the effects of filtering in the proposed DP-SCR and E-RescueDP, we also use the Kalman Filter (KF) to deal with the noise in DP-SCR.

Inspired by the FAST algorithm [17], KF [38] is used to improve the accuracy of the sanitized traffic flows $M(D^i_{t+1})$. The filtering mechanism includes two steps: *Predict* and *Correct*, which are shown in Algorithm 4.

Algorithm 4: Filtering with KF for $M(D^i_{t+1})$.

Require: the previous shared r^i_t and noisy measurement $Z^i_{t+1} = M(D^i_{t+1})$.
Ensure: the posterior estimate $\hat{x}^i_{t+1}$.
1: $KFPredict(t+1)$**:**
2: $\bar{x}^i_{t+1} = r^i_t$;
3: $\bar{P}^i_{t+1} = P^i_t + G$;
4: $KFCorrect(t+1)$**:**
5: $K^i_{t+1} = \bar{P}^i_{t+1}/(\bar{P}^i_{t+1} + H)$;
6: $\hat{x}^i_{t+1} = \bar{x}^i_{t+1} + K^i_{t+1}(z^i_{t+1}) - \bar{x}^i_{t+1}$;
7: $P^i_{t+1} = (1 - K^i_{t+1})\bar{P}^i_t$.

The posterior estimate $\hat{x}^i_{t+1}$ is the final shared traffic flow of section i at timestamp $t+1$, i.e., $r^i_{t+1} = \hat{x}^i_{t+1}$. The detailed principles and processes of KF have been explained in FAST algorithm, the readers may refer to [17].

Some experiments on filtering are conducted in this paper. As shown in Figure 5, the *MRE* of DP-SCR is slightly influenced by Kalman Filter and has a smaller value. It indicates that DP-SCR without Kalman Filter also has high accuracy. Thus, the sanitized traffic flows in DP-SCR without Kalman Filter can be shared directly.

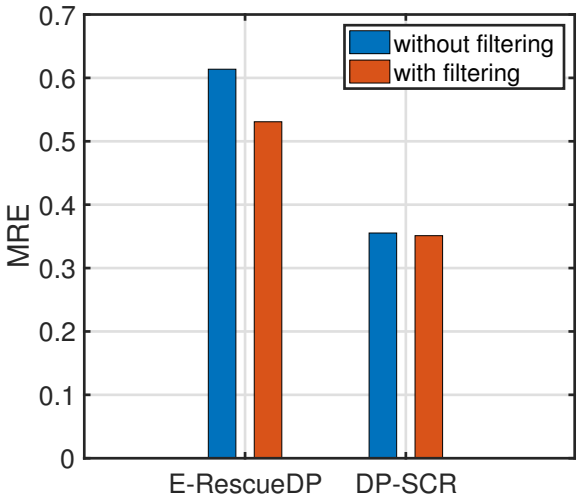

Figure 5. The effects of filtering for E-RescueDP and DP-SCR.

In addition, the *MRE* of E-RescueDP adopting and not adopting Kalman Filter are larger than that of DP-SCR. It indicates that DP-SCR is superior to E-RescueDP in terms of accuracy.

5.4. Complexity Analyses

The proposed DP-SCR scheme is compared with the other four schemes (i.e., BA, BD, E-RescueDP, and CLDP) in terms of time complexity, and the comparison results are shown in Table 2, where d is the number of sections, m is the number of groups/clusters in E-RescueDP and DP-SCR, and e represents the number of iterations required for the convergence of the 2-means in DP-SCR. As can be seen, BA, BD, and CLDP schemes are faster than E-RescueDP and DP-SCR, and DP-SCR may be faster than E-RescueDP when the number of sections is large.

Table 2. The comparison of complexity time.

schemes	BA [23]	BD [23]	E-RescueDP [24]	CLDP [29]	DP-SCR
complexity time	$\mathcal{O}(d)$	$\mathcal{O}(d)$	$\mathcal{O}(md^2)$	$\mathcal{O}(d)$	$\mathcal{O}(mde)$

6. Experimental Simulation and Evaluation

In this section, the related experiments are simulated on real-world datasets, and the performance of the proposed DP-SCR is compared with that of schemes E-RescueDP [24], BD [23] and BA [23]. All our experiments are run in Matlab 2018a platform on PC with Intel(R) Core(TM) i5-4590 CPU @ 3.30 GHz, 4.00 G main memory, and 500 GB hard disk with the Microsoft Windows 7 operating system.

The datasets of our experiments include the vehicular mobility dataset and the street layout dataset. The vehicular mobility dataset is mainly based on the real data collected by the General Departmental Council of Val de Marne (94) in France (Downloaded at http://vehicular-mobility-trace.github.io/ accessed on 2 March 2024). It comprises around 10,000 traces, over rush hour periods of two hours in the morning (7 a.m.–9 a.m.) and two hours in the evening (5 p.m.–7 p.m.). The real street layout of the Creteil roundabout area (sampled area) is obtained from the OpenStreetMap database, as shown in Figure 6. Here, each 400-m road is served as one section.

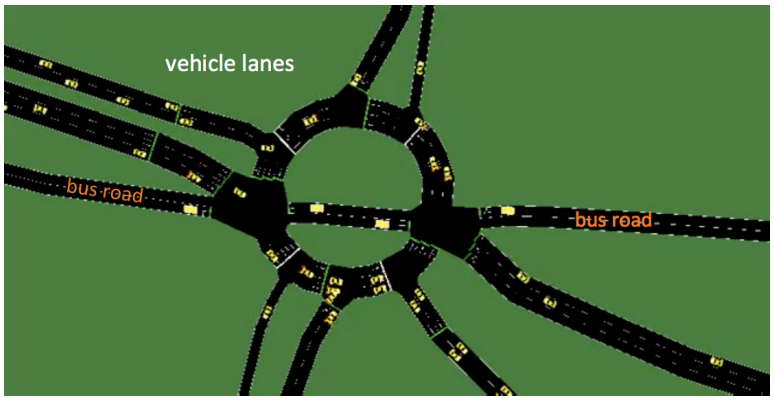

Figure 6. The street layout of training data.

Subsequently, a traffic flow dataset with 160 timestamps for each section is created, which is sampled every 85 s from the vehicular mobility dataset. Moreover, the traffic flow dataset contains vehicle numbers, vehicle coordinates on the two-dimensional plane (x and y coordinates in meters), vehicle speed (in meters per second), and vehicle id. The target section is randomly selected from the sections generated by function Q. In any case, to ensure the credibility of our experiments, all experiments involving the Laplace mechanism are conducted 100 times, and the average value of these 100 experiment results is represented by the points in the figures.

6.1. Data Utility of the Shared Traffic Flow

In this section, we conduct experiments for the designed adaptive allocation of privacy budget and dynamic clustering to evaluate the superiority in terms of data utility.

The accuracy of the shared traffic flows reflects the data utility. The mean absolute error (MAE) and the mean relative error (MRE) are served as an accuracy metric. Moreover, the smaller the MAE and MRE are, the more accurate the traffic flows are. Let $F^{i,n} = \{f^i_{1+m}, f^i_{2+m}, \cdots, f^i_{n+m}\}$ be raw traffic flows, and let $R^{i,n} = \{r^i_{1+m}, r^i_{2+m}, \cdots, r^i_{n+m}\}$ be the sanitized traffic flows of section i at successive n timestamps before Filtering. Then, the formulas for the MAE and MRE, respectively, are

$$MAE(F^{i,n}, R^{i,n}) = (1/n) \times \sum_{j=1}^{n} |f^i_{j+m} - r^i_{j+m}|, \text{ and}$$

$$MRE(F^{i,n}, R^{i,n}) = \frac{(1/n) \times \sum_{j=1}^{n} (|f^i_{j+m} - r^i_{j+m}|)}{\max(r^i_{j+m}, \delta)},$$

where δ is the bound of small traffic flows, which is used to reduce the effect of excessively small traffic flows and is equal to 0.1% of $\sum_{j=1}^{n} f^i_{j+m}$.

(1) Prediction accuracy evaluation for DP-SCR. The allocation for privacy budget affects the accuracy of the predicted traffic flows greatly. RescueDP and E-RescueDP are the baseline temporal-based schemes designed for the sharing of real-time data with w-event privacy. Due to the performance of E-RescueDP being better than that of RescueDP, we only compare our scheme with the preferable E-RescueDP in terms of the accuracy of prediction. In these experiments, the privacy budget is $\varepsilon = 1$.

E-RescueDP is based on the temporal correlation with the Elman network [39] (an RNN algorithm). In the Elman network, the number of neurons is 5 in the input layer, 18 in the hidden layer and 1 in the output layer, respectively. The designed diagram of the Elman network is shown in Figure 7. Moreover, the traffic flows at the first successive 80 timestamps of the target section is selected as the training sets, and the remaining traffic flows are taken as the testing sets. As depicted in Figure 8, the blue line represents the training loss (i.e., mean squared error) of the Elman network during training. It is noteworthy that the training loss remains stable and is equal to 0.012323 at 4997 epochs. In Figure 9, the blue dashed lines depict the raw traffic flows, while the orange solid lines represent the predicted traffic flows. Figure 9a displays the predicted results in the E-RescueDP scheme. The results show that there are significant differences between the predicted results of E-RescueDP and the raw traffic flows, where the MAR and MRE in E-RescueDP are 5.0875 and 0.4881, respectively.

In the proposed DP-SCR, the traffic data of the above target section and its linked sections at the last 80 timestamps are selected as experimental data. For fair comparison, the model M_v is trained on Elman network also with the number of neurons is 18 in the hidden layer and 1 in the output layer, respectively. The number of neurons is 3 in the input layer, including the shared traffic flow, maximal speed limit, and sampling period. The prediction results of DP-SCR are shown in Figure 9b, where the predicted traffic flows match the raw traffic flows well. Moreover, the MAR and MRE in DP-SCR are 3.1000 and 0.2680, respectively.

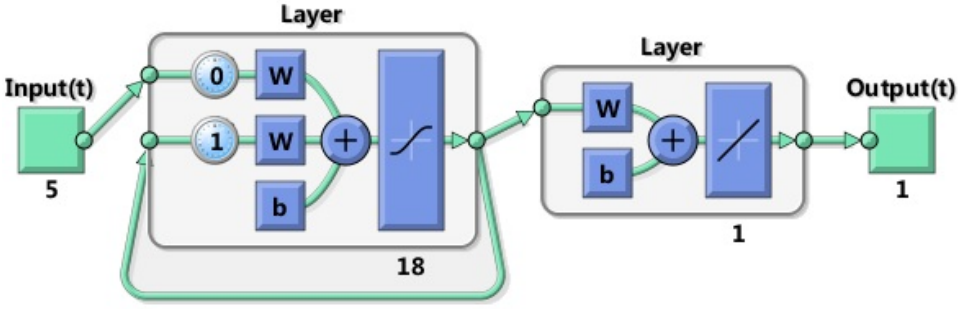

Figure 7. The designed diagram of the Elman network.

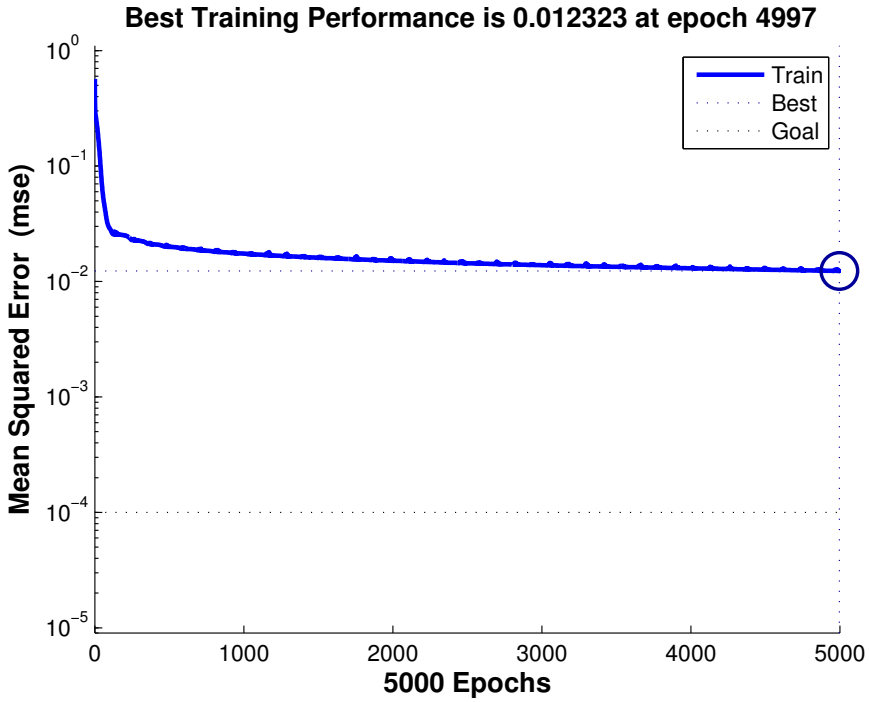

Figure 8. The training of the Elman network.

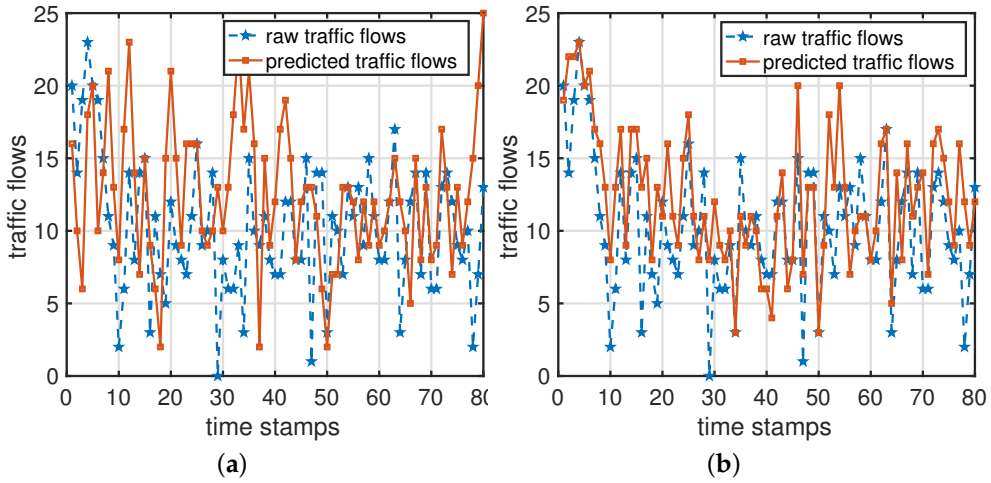

Figure 9. (**a**) The predicted results in E-RescueDP; (**b**) The predicted results in DP-SCR.

In summary, the predicted results in DP-SCR are more accurate than that in E-RescueDP, which means the data utility of the proposed scheme is higher.

(2) Accuracy evaluation for dynamic clustering. The dynamic clustering based on bisecting *k*-means is adopted to reduce the perturbation error caused by the small traffic flows, which will result in the loss of data utility. In the experiments on dynamic clustering, an experimental dataset based on real data is created, where the dataset includes 5000 sections with small traffic flows that are allocated with a random privacy budget. The *MRE* of the bisecting *k*-means is compared with that of non-partitioned operation and dynamic programming, where the dynamic programming is the partitioned operation in [24]. Figure 10 illustrates the *MRE* results of different strategies with different sections. It observes that the *MRE* of DP-SCR is smaller than that of other schemes, indicating the higher data utility of the shared traffic flows obtained by the dynamic clustering in DP-SCR compared to other partitioned operations.

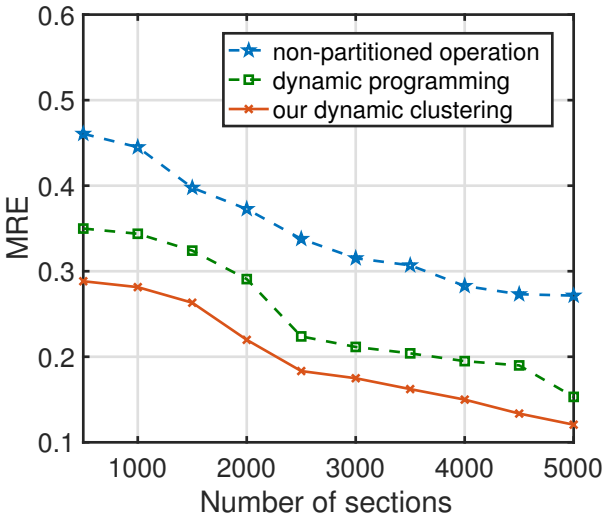

Figure 10. The effects of dynamic processing.

6.2. Data Utility vs. Privacy Budget (ε)

In this section, the experiments about the MAE and MRE are conducted when ε varies from 0.1 to 1.0. The MAE and MRE of DP-SCR are compared with those of schemes BA [23], BD [23], E-RescueDP [24] and CLDP [29]. BA and BD are baseline w-event privacy schemes for the real-time data release, and CLDP is the baseline w-event privacy scheme with local differential privacy. Figure 11 compares MAE and MRE for the shared traffic flows with ε changing, where w is fixed and equal to 10. The results indicate that the MAE and MRE of DP-SCR with any privacy budget are significantly smaller than those of other schemes. Moreover, the MAE and MRE of BA, BD and LCDP decrease as ε increases, and the magnitude of the decrease is also becoming smaller. The changes in the MAE and MRE of E-RescueDP and DP-SCR are little. There are three reasons for the above experimental results. First, BA, BD and CLDP allocate too small privacy budget for perturbation, which introduces more noise into the shared data. Second, the prediction of DP-SCR is more accurate than that of E-RescueDP, so the MAR and MER of DP-SCR are smaller than those of E-RescueDP. Third, since the adaptive allocation of budget privacy is adopted in E-RescueDP and DP-SCR, providing more reasonable privacy budgets, the MAR and MER of them are relatively stable when ε changes.

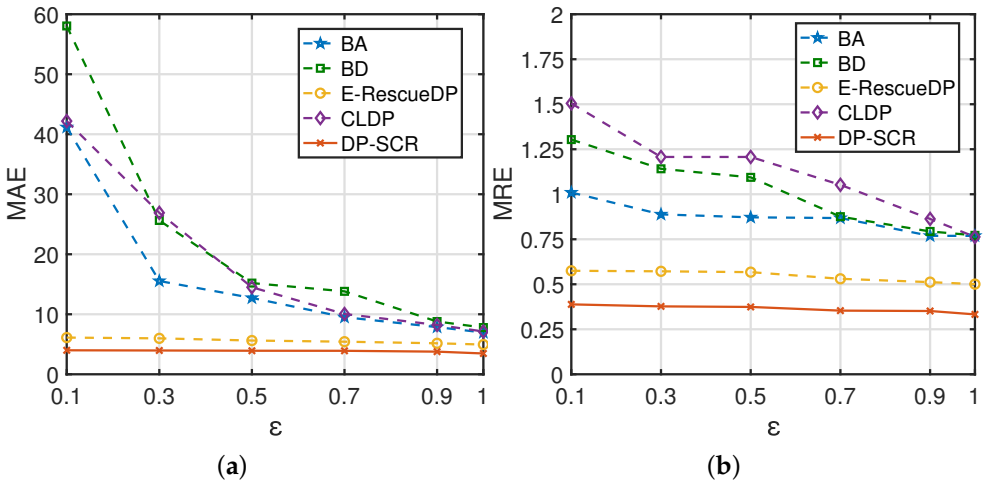

Figure 11. (**a**) The MAE of the shared traffic flows with ε changing (w = 10); (**b**) The MRE of the shared traffic flows with ε changing (w = 10).

6.3. Data Utility vs. Sliding Window Size (w)

The data utility of DP-SCR is compared with that of schemes BA, BD, E-RescueDP and LCDP, where w varies from 5 to 45. The results are shown in Figure 12, where the MAE and MRE of DP-SCR are higher than those of other schemes. Also, the MAE and MRE of BA, BD and LCDP increase as w increases, and the MAE and MRE of E-RescueDP and DP-SCR are relatively stable. This is because the adaptive allocation of privacy budget and dynamic processing improve the accuracy of the shared traffic flows and make them robust to the changes in w.

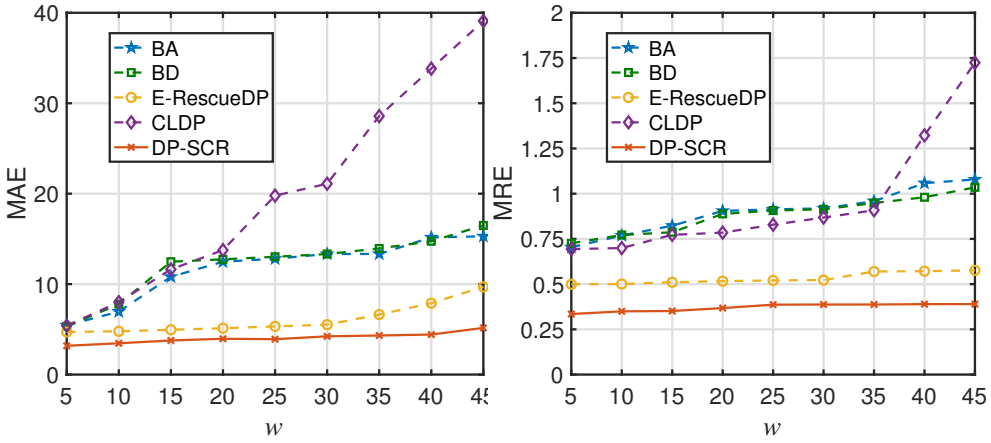

Figure 12. (**a**) The MAE of the shared traffic flows with w changing ($\varepsilon = 1$); (**b**) The MRE of the shared traffic flows with w changing ($\varepsilon = 1$).

7. Conclusions

In this paper, we propose a scheme, named DP-SCR, to ensure the sharing of real-time traffic flows with high data utility under privacy protection. DP-SCR consists of four key components: adaptive allocation of privacy budget, dynamic clustering, approximation and perturbation. In the proposed DP-SCR, we take advantage of the spatial correlation prediction and the novel clustering strategy to improve the accuracy of the shared traffic flows. Moreover, the results of the experiments on real-world datasets have also shown that the shared traffic flows in DP-SCR are more accurate than those in the existing baseline w-event privacy schemes. Also, in terms of privacy protection, DP-SCR has been proven to satisfy w-event privacy, which provides strong privacy protection to the shared traffic flows.

However, some aspects that still exist can be improved in future work. First, more characteristics of traffic flows may be considered together in prediction to improve the data utility. Second, genetic algorithms [40] may be used to improve the accuracy of the spatial correlation prediction. Finally, other privacy-preserving methods such as secure data deduplication [41], blockchain-based secure sharing scheme [42] and federated learning [43,44] can be used to enhance the data utility and security.

Author Contributions: Methodology, J.L., B.X., D.Z. and D.Q.; Software, J.L. and B.X.; Validation, J.L., B.X. and D.Q.; Formal analysis, J.L.; Writing—original draft, J.L. and D.Z.; Writing—review & editing, B.X. and D.Q.; Supervision, D.Z.; Funding acquisition, J.L. and D.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Natural Science Foundation of China (Grant no. 62202071, 62302072), in part by the China Postdoctoral Science Foundation (Grant no. 2022M710518, 2022M710520), and in part by the Natural Science Foundation of Chongqing, China (Grant no. CSTB2022NSCQ-MSX0358, CSTB2022NSCQ-MSX1217).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ding, X.; Zhou, W.; Sheng, S.; Bao, Z.; Choo, K.K.R.; Jin, H. Differentially private publication of streaming trajectory data. *Inf. Sci.* **2020**, *538*, 159–175. [CrossRef]
2. Li, L.; Jiang, R.; He, Z.; Chen, X.M.; Zhou, X. Trajectory data-based traffic flow studies: A revisit. *Transp. Res. Part Emerg. Technol.* **2020**, *114*, 225–240. [CrossRef]
3. Liu, Y.; James, J.; Kang, J.; Niyato, D.; Zhang, S. Privacy-preserving traffic flow prediction: A federated learning approach. *IEEE Internet Things J.* **2020**, *7*, 7751–7763. [CrossRef]
4. Le, J.; Lei, X.; Mu, N.; Zhang, H.; Zeng, K.; Liao, X. Federated Continuous Learning With Broad Network Architecture. *IEEE Trans. Cybern.* **2021**, *51*, 3874–3888. [CrossRef] [PubMed]
5. Yang, X.; Gu, B.; Zheng, B.; Ding, B.; Han, Y.; Yu, K. Toward Incentive-Compatible Vehicular Crowdsensing: An Edge-Assisted Hierarchical Framework. *IEEE Netw.* **2022**, *36*, 162–167. [CrossRef]
6. Chiou, J.M.; Liou, H.T.; Chen, W.H. Modeling time-varying variability and reliability of freeway travel time using functional principal component analysis. *IEEE Trans. Intell. Transp. Syst.* **2019**, *22*, 257–266. [CrossRef]
7. Wu, T.; Zhou, P.; Liu, K.; Yuan, Y.; Wang, X.; Huang, H.; Wu, D.O. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 8243–8256. [CrossRef]
8. Meese, C.; Chen, H.; Asif, S.A.; Li, W.; Shen, C.C.; Nejad, M. Bfrt: Blockchained federated learning for real-time traffic flow prediction. In Proceedings of the IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid), Taormina, Italy, 16–19 May 2022; pp. 317–326.
9. Miglani, A.; Kumar, N. Deep learning models for traffic flow prediction in autonomous vehicles: A review, solutions, and challenges. *Veh. Commun.* **2019**, *20*, 100184. [CrossRef]
10. Kiran, B.R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A.A.; Yogamani, S.; Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 4909–4926. [CrossRef]
11. Morlock, F.; Rolle, B.; Bauer, M.; Sawodny, O. Forecasts of electric vehicle energy consumption based on characteristic speed profiles and real-time traffic data. *IEEE Trans. Veh. Technol.* **2019**, *69*, 1404–1418. [CrossRef]
12. Gazdag, A.; Lestyán, S.; Remeli, M.; Ács, G.; Holczer, T.; Biczók, G. Privacy pitfalls of releasing in-vehicle network data. *Veh. Commun.* **2023**, *39*, 100565. [CrossRef]
13. De Montjoye, Y.A.; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. Unique in the Crowd: The privacy bounds of human mobility. *Sci. Rep.* **2013**, *3*, 1376. [CrossRef] [PubMed]
14. Dwork, C. Differential Privacy: A Survey of Results. In Proceedings of the International Conference on Theory and Applications of MODELS of Computation (TAMC), Xi'an, China, 25–29 April 2008; pp. 1–19.
15. Dwork, C.; Naor, M.; Pitassi, T.; Rothblum, G.N. Differential Privacy Under Continual Observation. In Proceedings of the Forty-Second ACM Symposium on Theory of Computing (STOC), Cambridge, MA, USA, 6–8 June 2010; pp. 715–724.
16. Chan, T.H.H.; Shi, E.; Song, D. Private and Continual Release of Statistics. *ACM Trans. Inf. Syst. Secur.* **2011**, *14*, 26:1–26:24. [CrossRef]
17. Fan, L.; Xiong, L. An Adaptive Approach to Real-Time Aggregate Monitoring With Differential Privacy. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 2094–2106.
18. Fan, L.; Xiong, L.; Sunderam, V. Differentially private multi-dimensional time series release for traffic monitoring. In *Differentially Private Multi-Dimensional Time Series Release for Traffic Monitoring*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 7964, pp. 33–48.
19. Chen, Y.; Machanavajjhala, A.; Hay, M.; Miklau, G. PeGaSus: Data-Adaptive Differentially Private Stream Processing. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), Dallas, TX, USA, 30 October–3 November 2017; pp. 1375–1388.
20. Ren, X.; Wang, S.; Yao, X.; Yu, C.M.; Yu, W.; Yang, X. Differentially Private Event Sequences Over Infinite Streams With Relaxed Privacy Guarantee. In *Differentially Private Event Sequences over Infinite Streams with Relaxed Privacy Guarantee*; Springer: Cham, Switzerland, 2019; Volume 11604, pp. 272–284.
21. Gati, N.J.; Yang, L.T.; Feng, J.; Nie, X.; Ren, Z.; Tarus, S.K. Differentially private data fusion and deep learning framework for cyber–physical–social systems: State-of-the-art and perspectives. *Inf. Fusion* **2021**, *76*, 298–314. [CrossRef]
22. Li, Q.; Heusdens, R.; Christensen, M.G. Communication efficient privacy-preserving distributed optimization using adaptive differential quantization. *Signal Process.* **2022**, *194*, 108456. [CrossRef]
23. Kellaris, G.; Papadopoulos, S.; Xiao, X.; Papadias, D. Differentially Private Event Sequences over Infinite Streams. *Proc. VLDB Endow.* **2014**, *7*, 1155–1166. [CrossRef]
24. Wang, Q.; Zhang, Y.; Lu, X.; Wang, Z.; Qin, Z.; Ren, K. Real-time and Spatio-temporal Crowd-sourced Social Network Data Publishing with Differential Privacy. *IEEE Trans. Dependable Secur. Comput.* **2016**, *15*, 591–606. [CrossRef]
25. Huo, Y.; Yong, C.; Lu, Y. Re-ADP: Real-Time Data Aggregation with Adaptive ω-Event Differential Privacy for Fog Computing. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 6285719. [CrossRef]
26. Wang, H.; Cai, S.; Liu, P.; Zhang, J.; Shen, Z.; Liu, K. DP-STGAT: Traffic statistics publishing with differential privacy and a spatial-temporal graph attention network. *Inf. Sci.* **2023**, *623*, 258–274. [CrossRef]
27. Wang, T.; Chen, J.Q.; Zhang, Z.; Su, D.; Cheng, Y.; Li, Z.; Li, N.; Jha, S. Continuous release of data streams under both centralized and local differential privacy. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), Virtual, 15–19 November 2021; pp. 1237–1253.

28. Ren, X.; Shi, L.; Yu, W.; Yang, S.; Zhao, C.; Xu, Z. LDP-IDS: Local differential privacy for infinite data streams. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Philadelphia, PA, USA, 12–17 June 2022; pp. 1064–1077.
29. Errounda, F.Z.; Liu, Y. Collective location statistics release with local differential privacy. *Future Gener. Comput. Syst.* **2021**, *124*, 174–186. [CrossRef]
30. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 265–284.
31. Dwork, C. Differential Privacy. In Proceedings of the International Conference on Automata, Languages and Programming (ICALP), Venice, Italy, 10–14 July 2006; pp. 1–12.
32. McSherry, F.D. Privacy Integrated Queries: An Extensible Platform for Privacy-preserving Data Analysis. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD), Providence, RI, USA, 29 June–2 July 2009; pp. 19–30.
33. Lu, H.; Sun, Z.; Qu, W. Big Data-Driven Based Real-Time Traffic Flow State Identification and Prediction. *Discret. Dyn. Nat. Soc.* **2015**, *2015*, 284906. [CrossRef]
34. Wang, W.; Li, W.; Ren, G. A speed-flow relationship model of highway traffic flow. *J. Harbin Inst. Technol.* **2005**, *12*, 331–335.
35. Alvarez-Marquez, A.; Aguilera, I.; Gentil, M.A.; Cabello, V.; Gonzalez-Escribano, M.F.; Nunez-Roldan, A. Traffic Flow Prediction for Road Transportation Networks With Limited Traffic Data. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 653–662.
36. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.Y. Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 865–873. [CrossRef]
37. Liebig, T.; Piatkowski, N.; Bockermann, C.; Morik, K. Dynamic route planning with real-time traffic predictions. *Inf. Syst.* **2017**, *64*, 258–265. [CrossRef]
38. Kalman, R. A new approach to linear filtering and predicted problems. *J. Basic Eng.* **1960**, *82*, 35–45. [CrossRef]
39. Elman, J.L. Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* **1991**, *7*, 195–225. [CrossRef]
40. Rangel, H.R.; Puig, V.; Farias, R.L.; Flores, J.J. Short-term demand forecast using a bank of neural network models trained using genetic algorithms for the optimal management of drinking water networks. *J. Hydroinform.* **2017**, *19*, 1–16. [CrossRef]
41. Zhang, D.; Le, J.; Mu, N.; Wu, J.; Liao, X. Secure and efficient data deduplication in jointcloud storage. *IEEE Trans. Cloud Comput.* **2021**, *11*, 156–167. [CrossRef]
42. Zhao, R.; Xu, C.; Zhu, Z.; Mo, W. A Blockchain-Based Secure Sharing Scheme for Electrical Impedance Tomography Data. *Mathematics* **2024**, *12*, 1120. [CrossRef]
43. Le, J.; Zhang, D.; Lei, X.; Jiao, L.; Zeng, K.; Liao, X. Privacy-preserving federated learning with malicious clients and honest-but-curious servers. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 4329–4344. [CrossRef]
44. Zhang, L.; Lei, X.; Shi, Y.; Huang, H.; Chen, C. Federated Learning for IoT Devices with Domain Generalization. *IEEE Internet Things J.* **2023**, *10*, 9622–9633. [CrossRef]