A REAL-TIME AGGREGATE DATA PUBLISHING SCHEME
WITH ADAPTIVE $\omega$-EVENT DIFFERENTIAL PRIVACY

CHENGTAO YONG AND YAN HUO*
School of Electronics and Information Engineering
Beijing Jiaotong University, Beijing, China

CHUNQIANG HU
School of Big Data & Software Engineering
Chongqing University, Chongqing, China

YANFEI LU AND GUANLIN JING
School of Electronics and Information Engineering
Beijing Jiaotong University, Beijing, China

(Communicated by Zhipeng Cai)

Abstract. Although massive real-time data collected from users can provide benefits to improve the quality of human daily lives, it is possible to expose users’ privacy. $\epsilon$-differential privacy is a notable model to provide strong privacy preserving in statistics. The existing works highlight $\omega$-event differential privacy with a fixed window size, which may not be suitable for many practical scenarios. In view of this issue, we explore a real-time scheme with adaptive $\omega$-event for differentially private time-series publishing (ADP) in this paper. In specific, we define a novel notion, Quality of Privacy (QoP) to measure both the utility of the released statistics and the performance of privacy preserving. According to this, we present an adaptive $\omega$-event differential privacy model that can provide privacy protection with higher accuracy and better privacy protection effect. In addition, we also design a smart grouping mechanism to improve the grouping performance, and then improve the availability of publishing statistics. Finally, comparing with the existing schemes, we exploit real-world and synthetic datasets to conduct several experiments to demonstrate the superior performance of the ADP scheme.

1. Introduction. Recently, context-aware smart devices are experiencing an unprecedented surge in smart cyber-physical networks [39]. These devices can achieve ubiquitous connectivity and powerful raw data-gathering capabilities via human-centric and device-to-device (D2D)-based applications [31]. Using data mining technologies, we then can discover knowledge from huge amounts of raw data. Accordingly, we learn that the technology may provides great benefits to push against our cognition of human behavior and to improve the quality of our lives.

2010 Mathematics Subject Classification. Primary: 58F15, 58F17; Secondary: 53C35.
Key words and phrases. Differential privacy, data publishing, privacy protection, data mining.

The first and second authors are supported by NSFC grants (No. 61471028) and the Fundamental Research Funds for the Central Universities (No. 2017JBM004).
The third author is supported by NSFC grants (No. 61702062).
The fourth author is supported by NSFC grants (No. 61571010).
* Corresponding author: Yan Huo.
Although this technology provides great benefits, sensitive and private information mined from raw data (e.g., social relationships and financial transactions) is also exposed to the risk of disclosure [6]. For example, more than 400,000 electronic eyes in Beijing may lead to privacy leakage (e.g., vehicle location information) by data sharing in vehicular ad hoc networks (VANETs) [18,30,32]. Similarly, we can also illegal access personal health datasets gathered from various sensors of physical sign in body sensor networks (BSNs) and publish these private data without permission [34]. As a result, data security has become an issue that cannot be ignored.

In general, data mining based threats may occur in three phases of data processing and transmission, i.e., cloud server attacks, data communication wiretap, and data inference [21]. The first two threats can be solved by cryptography based higher-layer security measures as well as emerging physical layer based secure transmit technologies. Comparing to the previous two attack modes, privacy leakage resulted from data analysis at servers is more serious for users. Therefore, our objective is to enable data publishers to share meaningful data in real-time while protecting their privacy.

The current state-of-the-art convincing model for providing privacy in data publishing is \( \epsilon \)-differential privacy [9,12]. The model requires perturbing the aggregate data prior to their publish [11]. In this model, statistics are collected and released by publishers and analyzed by data miners continually. In this way, we can protect sensitive information about the individuals that used for data mining based on the model. This model has been gradually applied to various streaming scenarios, e.g., disease surveillance [5] and real-time traffic monitoring [1]. These scenarios can be divided into two types protection modes to address the issue of real-time data publishing with \( \epsilon \)-differential privacy guarantee: event-level privacy protection and user-level privacy protection [14]. The former one may protect a single event occurred by users at specific timestamps while the latter may hide all the events of users throughout the entire stream. So far, most of the state-of-the-art strategies focuses on either event-level privacy model over infinite streams [23,24] or user-level privacy model over finite streams [17].

Considering the above two privacy protection models, Kellaris et al. propose an \( \omega \)-event \( \epsilon \)-differential privacy model in [26] to analyze and compare their inherent strengths and weaknesses. For infinite data streams, Wang et al. [36] design a scheme to accomplish \( \omega \)-event privacy model. They employ a fixed size sliding window via ignoring time series changes, which should not be suitable for the scenario of changeable and irregular data. In addition, their scheme also introduce large errors in the grouping and perturbation mechanism due to the use of average operation. To solve these challenges, we present a real-time aggregate publishing framework with adaptive \( \omega \)-event differential privacy (ADP). The ADP framework is composed of a quality of privacy (QoP)-based adaptive \( \omega \)-event mechanism, a smart grouping-based perturbation mechanism, and a filtering mechanism. Using these mechanisms in cooperation, the proposed integrated framework can protect privacy of publishing statistics over infinite timestamps.

The main contributions of the paper are summarized as follows. Taking the window size \( \omega \) and errors of the published statistics into consideration, we pioneer a novel metric, i.e., quality of privacy (QoP). Using this metric, we intend to adjust the size of window \( \omega \) adaptively by dint of the design of QoP-based adaptive \( \omega \)-event mechanism. Next, in order to improve the grouping effect when achieve
differential privacy, we exploit a machine learning based technology to design a smart grouping scheme. Moreover, considering the high correlation among each event in the temporal dimension, we utilize the Kalman filter to reduce errors of released data. Finally, we prove that our proposed ADP can satisfy differential privacy and demonstrate its high utility and security through a large number of practical experiments.

The rest of the paper is organized as follows. In Section 2, we review previous works related to differential privacy on static data and streams. The whole ADP framework and the corresponding common applications are illustrated in Section 3. Next, we present QoP-based adaptive \( \omega \)-event privacy algorithm in Section 4, including a dynamical adjustment method of the window size \( \omega \). Moreover, Section 5 introduces a smart grouping-based perturbation algorithm, which can reduce the noise added to data significantly. In Section 6, we analyze whether the ADP framework meets differential privacy and then provide several simulation results to discuss the performance of each mechanism in our framework. Finally, Section 7 concludes our paper and states promising directions for future work.

2. Related work. In this section, we survey the latest studies about differential privacy on both static data and data streams briefly.

2.1. Differential privacy on static data. For privacy of static data, Dwork et al. first proposed a standard method in [16], which introduced the Laplace noise into aggregate queries to guarantee differential privacy.

Ever since, various methods have been designed to enhance the accuracy of released data with extra noise. In [4], Blum et al. demonstrated that privacy-preserving databases could be released for all queries in a discrete domain but not in continuous domains. Then, in order to release useful information for halfspace queries, they designed a privacy-preserving polynomial time algorithm. Dwork et al. further proposed more efficient algorithms in [15] to release private sanitization of a data set with hardness results obtained. The work of Hay et al. exploited consistency check to improve the accuracy of a tree of counting queries. Specifically, they introduced constrained inference techniques to transform the differentially private output into a consistent one in a post-processing procedure. This hierarchical structure of queries was finally referred to as a histogram mode, e.g., in [22, 28, 33, 38], where each level in the tree is an increasingly fine-grained summary of data. In a recent work [37], the authors designed a scheme that injected different amount of the Laplace noise to different query results. In this case, we could employ correlated noises to improve data utility. Furthermore, several other works [2, 8, 10, 20, 29] investigated different protection schemes for particular kinds of data with differential privacy, such as search logs, sparse data and set-valued data.

According to the above analysis, the above methods were all designed to perturb static data, so they are not applicable for time-series data with high self-correlation.

2.2. Differential privacy on streams. Considering privacy of data streams, Dwork et al. in [14] improved upon the counter publishing schemes of [13]. They focused on a stream with the fixed length \( T \) and built a full binary tree over stream update. Every node in the tree stored noise values with scale logarithmic in \( T \). Then, at the \( i \)th update, these nodes identified the subtrees they belonged to and reported the current counter after adding the noise values stored in the roots of these subtrees. Similarly, Chan et al. proposed an \( \epsilon \)-differentially private continual
counter with small errors in [23]. Specifically, they guaranteed $O(\frac{1}{\epsilon} \cdot (\log t)^{1.5} \cdot \log \frac{1}{\sigma})$ errors for each $t \in \mathbb{N}$. Moreover, the great improvement in [23] was that their mechanism did not require a priori knowledge of an upper bound on the time for which it ran and provided guarantee even when it ran indefinitely.

Different from the counter-based setting, Cao et al. in [7] studied a protection method for sensitive streams within a window instead of the whole infinite stream. Considering window-based applications, they explored a stream-based management system to cope with numerous aggregate queries simultaneously. In [17], Fan and Xiong intended to hide all events of users and designed a user-level privacy strategy for a finite stream. For the received perturbing data, they employed the Kalman filter [3] in their strategy to improve accuracy of differentially private data release.

Considering multiple events occurring at continuous time segments, Kellaris et al. presented an $\omega$-event $\epsilon$-differential privacy model in [26]. This model combined the advantages of event-level privacy model and user-level privacy model skillfully. In the model, they employed a sliding window to capture a wide range of $\omega$-event privacy, and designed a scheme to distribute and absorb the privacy budget on the assumption of not significantly changeable statistics. On this basis, Wang and Zhang further designed an online aggregate monitoring scheme (called as RescueDP) for infinite streams in [36]. Their scheme integrated adaptive sampling and budget mechanism as well as dynamic grouping and perturbation to provide privacy-preserving statistics.

Despite the ongoing research having played a vital role for differential privacy on streams, there still exist challenges to be handled. We point out that, the fixed sliding windows employed in most existing frameworks may not be practical. In addition, existing metrics are for static data, but very little for streaming media. Motivated by these challenge, in this paper, we present adaptive $\omega$-event $\epsilon$-differential privacy model and quality of privacy (QoP).

3. The framework of adaptive $\omega$-event differential privacy. In this paper, we consider a universal scenario where the statistics (e.g., trajectories of vehicles and personal health data) of each region are published continually. Fig. 1 explains a trustful cloud server that gathers data from a large number of data contributors and publishes the sanitized data to a data mining-enabled system. In the model, we assume that the server and data transmission links are secure absolutely. In other words, the only possibility to disclose sensitive information is published data in the process of data mining [40]. Thus, the goal of our research is enable these data contributors to share useful information continuously under the premise of protecting their privacy.

For the goal of our research, we propose a privacy preserving solution based on adaptive $\omega$-event differential privacy (ADP). This solution, shown in Fig. 2, consists of three parts. For the first part, we employ adaptive sampling and the result of QoP to achieve the QoP-based adaptive $\omega$-event privacy mechanism. Then, the smart grouping-based perturbation is designed by the smart grouping and perturbation mechanism. Moreover, we also adopt the FAST mechanism in [17] to reduce errors of published data.

For the Fig. 2, we provide the detailed procedures of the proposed ADP. At first, we propose quality of privacy (QoP) to measure privacy quality, with the consideration of window size $\omega$ and errors of published statistics. By using QoP as metric, we can obtain the optimal number of sampling points in successive $\omega$. 
timestamps. Besides, we also employ adaptive sampling to adjust the sampling interval dynamically. These steps are described in the red dashed box in the Fig. 2. In these steps, the private budget will not be allocated to non-sampled points and the statistics of non-sampled points is approached by sampled points. As a result, the size of the Laplace noise adding to row statistics could be reduced while the utility of the released statistics could be increased. Finally, the window size can be adjusted adaptively due to artificially determined number of samples and sampling interval. In other words, the size of windows can be reduced when data increases rapidly, and vice versa.

For the RescueDP strategy in [36], the grouping mechanism may destroy the spatial correlation of regions due to lack of the consideration of the overall data changes. Besides, the mechanism may be not enough effective with a fixed threshold in most practical scenarios. Therefore, we design a new smart grouping mechanism to aggregate regions based on the properties of predicted data, which is shown in the blue dashed box in the Fig. 2. All in all, we introduce the auto-regressive integrated moving average model (ARIMA) into the smart grouping to dynamically predict predict statistics of current timestamp on the basis of historical statistics. In addition, the affinity propagation (AP) algorithm [19] is exploited to aggregate
regions smartly. Then, a Laplace mechanism is applied to perturb true values of statistics base on proposed grouping mechanism.

Here, we first elaborate the complete procedures of the ADP strategy in Algorithm 1. Then the detailed design of every components is illustrated in next sections.

Algorithm 1 The Adaptive \(\omega\)-event Differential Privacy

**Input:** A predefined raw database \(X_i\) at the timestamp \(i\)

**Output:** Released statistics \(R_i\)

1: Find out the optimal number of sampling points \(N\) with the use of QoP as a measure standard
2: Find out the set of sampling regions at the current timestamp
3: Obtain the grouping strategy \(G_{ki}\) with the smart grouping mechanism on sampling regions
4: Allocate budgets for all sampling regions
5: Introduce the Laplace noises into the group \(G_{ki}\) with the allocated budget at the perturbation mechanism
6: Employ the filtering mechanism to improve the accuracy of perturbed statistics and release sanitized statistics \(R_i\).
7: Update the sampling interval using the adaptive sampling mechanism

4. **QoP-based adaptive \(\omega\)-event privacy.** For privacy protection on infinite streams publishing, \(\omega\)-event privacy is a convincing model. The objective is to make a trade off between utility and privacy to protect all event sequences that occur within all windows of \(\omega\) timestamps. However, it is not applicable to many realistic scenarios due to the fixed size of the sliding window. The key issue of the unrealistic assumption is that most real-time aggregate data streams generated from monitoring application are significantly different in various time periods. For example, within successive timestamps, it can be seen that traffic data varies sharply in the daytime but is relatively stable at night. Thus, we introduce a new QoP-based adaptive \(\omega\)-event privacy mechanism in this section to dynamically adjust the window size \(\omega\) within different timestamps. Following three subsections describe the key parts to achieve this mechanism, including the QoP definition, the adaptive sampling design and the adaptive \(\omega\)-event privacy design.

4.1. **Quality of privacy.** Considering the window size \(\omega\) and errors of published statistics, QoP is proposed to measure the corresponding privacy quality. Assume \(x = \{x_1, \ldots, x_k\}\) and \(r = \{r_1, \ldots, r_k\}\) represent the raw time series in a window and the sanitized time series, respectively. Then, we exploit mean absolute error (MAE) to measure difference between these two time series.

\[
MAE(x, r) = \frac{1}{k} \sum_{i=1}^{k} |r_i - x_i|.
\]  

(1)

Yet, we employ a sampling mechanism in the proposed ADP. It may perturb statistics at selected timestamps and approximate the non-sampled statistics with perturbed sampled statistics. Thus, (1) can be rewritten as follows.

\[
MAE(x, r) = \frac{1}{k} \sum_{i}^{k} \sum_{j=i}^{i-1} |r_i - x_j|.
\]  

(2)
As a result, QoP in a window is defined as
\[ QoP(x, r) = \theta \frac{\omega}{MAE(x, r)}, \]  
where \( \omega \) is a window size and \( \theta \) is a pre-defined scale factor to adjust the weight between \( \omega \) and \( MAE(x, r) \). Intuitively, as the events occurred in contiguous time-stamps are not independent, there is close correlation among these events when data changes slowly. Meanwhile, with the possibility that sensitive information may be disclosed, the window size \( \omega \) should be increased when data changes slowly.

4.2. The adaptive sampling design. In general, the release of noisy data denotes the expenditure of fixed budget \( \epsilon \). If the window size \( \omega \) is large, the publishing of noisy data will emerge gigantic errors. This problem can be addressed by using a sampling mechanism. This mechanism can perturb sampled statistics while skipping non-sampled statistics\(^1\). In this case, we can employ skipping a point to save budget for future perturbation.

Inspired by the RescueDP strategy, we exploit a proportional-integral-derivative (PID) controller to change the sampling rate based on dynamic historical data. In the RescueDP algorithm, the authors uses non-fixed sampling points but may result in the fact that the budget may not be fully utilized. Different from the RescueDP algorithm, we adopt the optimal number of sampling points. As a result, the budget allocation strategy is equal to each of the sampling points. The measurement of feedback errors in the region \( j \) is defined as follows.
\[ E_{j k_n}^j = |r_{j k_n}^j - r_{j k_n-1}^j|, \]  
where \( k_n \) and \( k_{n-1} \) indicate the current sampling point and the last sampling point, respectively. It shows that data changes rapidly when the error \( E_{j k_n}^j \) increases. Then, the full PID error \( \delta^j \) of region \( j \) is defined as below.
\[ \delta^j = C_p E_{k_n}^j + C_i \frac{\sum_{o=n-\pi}^{n-1} E_{k_o}^j}{\pi} + C_d \frac{E_{k_n}^j - E_{k_n-1}^j}{k_n - k_{n-1}}, \]  
where \( C_p \), \( C_i \), and \( C_d \) denote proportional gain, integral gain and derivative gain, respectively.

Intuitively, the sampling interval needs to be small with rapid data change. Thus, a new sampling interval \( I \) is defined as follows.
\[ I = \max\{1, I_l + \theta(1 - (\delta^j)^2)\}, \]  
In (6), \( I \) and \( I_l \) denote the existing and the last sampling interval of region \( j \), respectively. Unlike the RescueDP strategy, the PID error is the only factor that influences the sampling interval. The reason is that the budget is equally allocated at all sampling points. In other words, Therefore, data dynamics in the adaptive sampling design can be better characterized by the sampling interval.

4.3. The adaptive \( \omega \)-event privacy design. Suppose that the starting and ending points of the window are both sampling points, the size of window \( \omega \) is determined as follows,
\[ \omega = \sum_{k=1}^{n-1} I_k, \]  
\(^1\)The non-sampling statistics can be approximated by the last published statistics.
where \( n \) is the number of sampling points in current window. \( I_k \) represents the sampling interval. For example, \( I_0 \) is the interval between the first and the second sampling point in current window. Obviously, the size of the window is determined by \( n \) and \( I_k \). \( I_k \) is adaptive adjust by Adaptive Sampling Mechanism and \( n \) can be obtained as follows.

According to (1), (3), and (7), QoP in a window can be calculated as follows.

\[
QoP(x, r) = \theta \frac{\sum_{k=1}^{n-1} I_k}{\frac{1}{k} \sum_{i} \sum_{j=1}^{i-1} |r_i - x_j|},
\]

Then the optimal number of sampling points \( N \) that maximizes \( QoP \) is defined as

\[
\max \ QoP \quad \text{s.t.} \quad 2 \leq N \leq \omega,
\]

where \( \omega \) is the mean of all window size over all timestamps. As a result, the optimal number of sampling points \( N \) can be obtained if the sampling interval \( I_k \) and perturbed statistics is known. After obtaining \( N \), we enable \( n \) identical to \( N \).

Next, the procedures of the adaptive \( \omega \)-event privacy mechanism are described as follows. Note that \( N \) can be obtained by machine learning from history data and \( I_k \) can be obtained by an adaptive sampling mechanism. Assume that the next sampling point is \( I \) and \( I_0, \cdots, I_{N-1} \) are sampling intervals in the current window, respectively. When the window slides to the next sampling point, we first observe the number of sampling point \( n \). If \( N > n \), we should increase the size of the window when \( I > I_0 \), i.e.,

\[
\omega = \omega_l + \Delta,
\]

where \( \omega \) and \( \omega_l \) are the next and last window, respectively. \( \Delta \) is the difference between \( I \) and \( I_0 \). Here, \( N \geq n \) holds because the ending point of the window can slide to the next sampling point while the starting point of the window cannot slide to the second point when \( I < I_0 \). In this case, we should decrease the window size by moving the starting point of the window to the second point. In addition, \( \omega_l \) is not changeable when \( I = I_0 \).

5. Smart grouping-based perturbation.

5.1. The smart grouping algorithm. A naive method to achieve differential privacy is to inject the Laplace noises to statistics. Nonetheless, it is likely to introduce more perturbation errors, especially in statistics with small values. Therefore, the authors in [36] present a dynamic grouping algorithm to aggregate regions with small statistics together in a dynamic way with the change of statistics. Yet, dynamic grouping algorithms also show limitations in some aspects. First, the method to predict statistics at the sampling point \( t_i \) in region \( j \) may not be accurate, which can lead to greater errors. Second, it may not be suitable to employ fixed thresholds of \( \tau_1 \), \( \tau_2 \), and \( \tau_3 \) to achieve dynamic grouping in realistic scenarios. Therefore, we present a smart grouping algorithm in this subsection.

The proposed grouping algorithm can smartly aggregate regions with small statistics in the noise scenarios. Different from the RescueDP strategy, our grouping algorithm is mainly based on machine learning algorithm. Thus, the ARIMA model is firstly used to predict statistics at the existing sampling point \( t_i \) in region \( j \), i.e., \( \hat{r}_i^j \). Then, we just use one threshold \( \tau \) that can be dynamically generated according
to the budget allocated to each sampling point $\epsilon_i$. It is shown in this research that
the inverse proportion could better represent the relationship between the anti-noise
threshold $\tau$ and the allocation budget for each timestamp $\epsilon_i$. At last, we implement
the affinity propagation (AP) algorithm [19] to aggregate regions, where the predict
statistics $\hat{r}_{ij}^t$ is less than the anti-noise threshold $\tau$.

The smart grouping algorithm is presented in Algorithm 2. It can be seen
that only the regions that need to be sampled (denoted by $S$) can be grouped at
timestamps. We can divide Algorithm 2 into four parts. The first one called as
region filtering is from line 1 to line 7, which can filter out the candidate regions
based on $\hat{r}_{ij}^t > \tau$. Next, we initialize three parameters that are used to represent
data characteristics, i.e., similarity $s(i, j) = -\|\hat{r}_{ij}^t - \hat{r}_{ik}^t\|^2$, responsibility $r(i, j)$, and
availability $a(i, j)^2$. In the third step (from line 11 to line 16), we can find the
cluster center $x_i$ for each region $i$ and determine the corresponding clusters $C_{x_i}$
when $r(i, j)$ and $a(i, k)$ converge (i.e., do not change significantly within a fixed
number of iterations). Finally, we can achieve the grouping results through line 17
of Algorithm 2.

Algorithm 2 Smart Grouping Algorithm

Input: $S_{t_i}$: the regions that need to be sampled at $t_i$

Output: $G_{t_i}$: The grouping strategy at $t_i$

1: for each region $i$ in $S_{t_i}$ do
2: Predict statistics (denoted by $\hat{r}_{ij}^t$) using the ARIMA model
3: if $\hat{r}_{ij}^t > \tau$ then
4: Let the region $i$ itself as a group
5: Add the group to $G_{t_i}$
6: else
7: Add the region into $\Phi$
8: Calculate the similarities $s(i, j)$ between pairs of regions in $\Phi$;
9: Initialize all $a(i, j)$ to be zero;
10: Initialize $r(i, j) = \frac{\sum_{i,j} s(i, j)}{N^2}$ ;
11: while not convergent do
12: $r(i, k) \leftarrow s(i, k) - \max_{k' \text{ s.t. } k \neq k'} \{a(i, k') + s(i, k')\}$
13: $a(i, k) \leftarrow \left\{ \begin{array}{ll}
\sum_{i', j'} \max\{0, r(i', k)\}, & \text{for } k = i \\
\min\{0, r(k, k) + \sum_{i', j' \notin \{i, k\}} \max\{0, r(i', k)\}\}, & \text{for } k \neq i
\end{array} \right.$
14: Cluster centers: $F = \{x_i | x_i = \arg \max_k \{(a(i, k) + r(i, k))\}$
15: Compute all $C_{x_i} = \{j | x_j = x_i\}
16: Cluster result: $P = \{C_j | j \text{ is a cluster center}\$
17: Group the regions in $\Phi$ according $P$ and add each group to $G_{k_i}$
18: return The grouping strategy $G_{t_i}$

Here, $r(i, k)$ denotes the accumulated evidence for how well-suited point $k$ is to serve as the
exemplar for point $i$ while $a(i, k)$ represents the accumulated evidence for how appropriate it would be
for point $i$ to choose point $k$ as its exemplar.
5.2. Smart grouping-based perturbation. Based on the design of adaptive sampling, we intend to deliberately inject additional noises into publishing statistics by using the Laplace mechanism [16]. These publishing statistics do not include the non-sampled statistics that can be approximated by the last publishing statistics.

To achieve the additional noises loading, we present a method of smart grouping-based perturbation. This method consists of a perturbation step and an allocation step. In the perturbation step, the Laplace mechanism is applied in each group rather than in each region due to the utilization of the grouping algorithm. Here, we first provide the definition of the Laplace mechanism as follows.

**Definition 5.1.** (The Laplace mechanism [16]) Let \( D \) denote a set of datasets, for any function \( f: D \rightarrow \mathbb{R} \), the Laplace Mechanism \( f \) for any dataset \( D \in D \)

\[
R(D) = f(D) + \text{Lap}(\Delta(f)/\epsilon),
\]

where the noise follow a Laplace distribution with mean zero and scale \( \Delta(f)/\epsilon \). And the \( \Delta(f) \) is sensitivity of \( f \), which defined as the maximum \( L_1 \) norm for all neighboring dataset \( D, D' \in D \).

The RescueDP strategy may not make full use of the total budget. The reason is that it takes the smallest budget of a region in a group as the budget of the entire group for the purpose of avoiding exceeding the total budget. Different from this, we fix the number of sampling points and allocate the total budget to each sampling point uniformly in our algorithm. As a result, the budget used for a group is the sum of the budget of each region. The total budget can be utilized fully.

We assume that a group \( g \) has \( \kappa \) regions, i.e., \( g \) is composed of \( \kappa \) columns of \( D_i \) \( (g \subseteq D_i) \). In addition, \( f(g) \) represents a function to aggregate the number of data contributors in \( g \). Intuitively, because all contributors can only appear in one region at one timestamp, the sensitivity of the function \( f \) is equal to 1, i.e., \( \Delta(f) = 1 \). Then, the Laplace mechanism can be employed in group \( g \) as follows,

\[
R(g) = f(g) + \text{Lap}(\lambda(g)) = \sum_{j=1}^{\kappa} g[j] + \text{Lap}(\Delta(f)/\epsilon_i),
\]

where \( g[j] \) is the \( j \)th region of the group \( g \) and \( \lambda(g) \) represents the scale of Laplace noises injected into \( f(g) \).

In the allocation step, we allocate the group perturbed statistic to each region according to the predicted statistics of each region. In this way, we can avoid errors resulted from the average operation adopted in the RescueDP strategy. Our allocation method is shown as follows.

\[
R(g[j]) = \alpha_j M(g), \quad \forall j = 1, ..., \kappa,
\]

where \( \alpha_j \) is the weight of each region. It can be calculated by the predicted statistics of each region, i.e.,

\[
\alpha_j = \frac{\hat{r}_j}{\sum_{j=1}^{\kappa} F_j}.
\]

In this way, the perturbed statistics of each region is more accurate.

6. Performance discussion. In this section, we first analyze the privacy of our proposed ADP framework in theory, then provide several numerical simulation results to study the performance of our framework in terms of MAE as well as QoP.

\[3\text{The neighboring dataset is different on at most one record.}\]
6.1. Privacy analysis.

**Theorem 6.1.** The proposed ADP framework satisfies $\epsilon$-differential privacy.

**Proof.** In the ADP framework, perturbation is the only possible mechanism to disclose private information because it is the only one to access raw data. As a result, if we prove that the perturbation mechanism satisfies $\epsilon$-differential privacy, the ADP framework can meet the requirement of $\epsilon$-differential privacy subsequently.

On the basis of the smart grouping strategy $G_t$, each group includes several disjoint regions. We assume that $g_i$ with $\kappa_i$ regions is an arbitrary group of $G_t$. According to (8), the laplace mechanism on group $g_i$ is as follows,

$$R(g_i) = f(g_i) + \text{Lap}(\lambda(g_i)) = \sum_{j=1}^{\kappa_i} g_i[j] + \text{Lap}(\Delta(f)/\epsilon_i),$$  \hspace{1cm} (15)

where $g_i[j]$ is the $j$th region of $g_i$ and $\Delta(f) = 1$.

Based on Definition 5.1, $R(g_i)$ satisfies $\epsilon_i$-differential privacy. According to Axiom 2.1.1 in [27], post-processing sanitized data will not reveal privacy as long as the sensitive information is not available directly in the post-processing algorithm. As a result, $R(g_i[j]), \forall j = 1, \cdots, \kappa_i$, also satisfies $\epsilon_i$-differential privacy. Assume that $\epsilon'_i$ and $\epsilon_i$ represent the budget used for perturbation and the budget by the budget allocation mechanism for a region at timestamp $i$, respectively. If all allocated budget is employed for perturbation in our algorithm, then $\epsilon'_i = \epsilon_i$ holds.

Based on Theorem 3 in [26], the perturbation mechanism of a region satisfies $\epsilon$-differential privacy for every $t$ and $k \in [t]$, if it holds that

$$\sum_{i=k-\omega+1}^{k} \epsilon'_i \leq \epsilon.$$  \hspace{1cm} (16)

This means that the budget allocation mechanism can ensure $\sum_{i=k-\omega+1}^{k} \epsilon'_i \leq \epsilon$ for any sliding window $\omega$ timestamps for the reason that $\epsilon'_i = \epsilon_i$ holds. Thus, the perturbation mechanism on each group can satisfy $\epsilon$-differential privacy. In other words, the ADP algorithm also satisfies $\epsilon$-differential privacy.

6.2. Numerical simulation results. We compare the performance of our ADP algorithm with the RescueDP strategy and the BA scheme over real and synthetic data sets. The RescueDP and the BA are two latest schemes that provide $\epsilon$-event privacy for real time aggregate data publishing. We employ MAE and QoP as metrics to study the performance of the three schemes. The specific calculation of these metrics is given by (1) and (8). Our experiment is conducted in Python environment in Windows 7 operating system.

In our experiment, we use two real-world data sets, i.e., Capital [42] and CTR [41], to discuss the performance. The dataset of Capital provides an accurate data containing the bikeshare trip in Washington D.C. for one year from January 1, 2016 to December 31, 2016. It contains a total of 3,333,791 bikeshare trajectories. Each of them consists of the bike number, the end station and time, and the start station and time. The dataset of CTR consists of ad click records between October 21, 2014 and October 30, 2014. It contains 40,428,967 records and each record includes an ad ID, a device IP and a timestamp, etc. And then, the number of clicks of each ad is published per hour. In addition, we also exploit a famous moving users generator, named as San, by Brinkhoff in [35] to generate a synthetic dataset. Also, the map
of San Joaquin is also used in our experiments. More detailed settings can refer to [36].

6.2.1. Utility v.s. privacy. In Fig. 3, we compare our ADP scheme, the RescueDP scheme, and the BA scheme in terms of the trade-off between utility and privacy. It is clear that when $\epsilon$ increases, the QoP of three schemes grow gradually while the MAE decreases. The reason is that the larger $\epsilon$ means the smaller noise need to be injected. Moreover, for three data sets, the ADP scheme outperforms the RescueDP scheme and the BA scheme greatly, especially in a small privacy budget.

![Graphs showing utility and privacy comparison](image)

**Figure 3.** Utility comparison when $\epsilon$ changes.

The superior performance of the ADP scheme results from the following three aspects. First, due to the design of the optimal number of sampling points and the corresponding privacy budget allocation mechanism, the privacy budget is fully used for private perturbation. Second, the adaptive $\omega$-event privacy mechanism in the ADP scheme satisfies the privacy window adaptively and improves the practicability of the scheme. Finally, the smart grouping-based perturbation mechanism can provide better grouping results to reduce the error introduced by Laplace mechanism.

6.2.2. Effect of adaptive $\omega$-event privacy mechanism. In order to highlights the advantages of adaptive $\omega$-event privacy mechanism, we compare our ADP scheme with a variant version, $ADP(f)$, which only adapts fixed $\omega$-event privacy mechanism. Fig. 4 demonstrates the comparison results in terms of MAE and QoP. It can be clearly seen that the adaptive $\omega$ mechanism can increase QoP while decrease MAE significantly in both real-world and synthetic datasets. Therefore, we can draw the conclusion that the adaptive $\omega$-event privacy mechanism advances the quality of release data considerably and it is more down-to-earth for many real-world scenarios.
6.2.3. Effect of smart grouping mechanism. In this part, we investigate the performance of our smart grouping mechanism as well as the traditional grouping mechanism proposed in the RescueDP scheme [36]. As shown in Fig. 5, both MAE and QoP of the smart grouping mechanism is exceed the traditional grouping mechanism. The excellent performance of smart grouping is chiefly benefit from the application of machine learning algorithms.

7. Conclusion. In this paper, we present a real-time scheme with adaptive \(\omega\)-event for differentially private time-series publishing (ADP). This scheme is composed of a QoP-based adaptive \(\omega\)-event privacy mechanism, a smart grouping-based perturbation mechanism, and a filtering mechanism. The key innovation is that we propose adaptive \(\omega\)-event privacy that can adjust the size of the protection window adaptively. It can be more adapted to practical scenarios. In order to evaluate privacy quality more comprehensively, we present QoP with the protection window size. Moreover, an affinity propagation (AP)-based new grouping mechanism is designed to make accurate grouping. Finally, we provide a theory to prove that ADP satisfies differential privacy in theoretical. Extensive experiments over real-world and
synthetic datasets show that the ADP scheme outperforms existing methods and improves the utility of real-time data publishing with strong privacy preserving.

Acknowledgments. We are very grateful to all reviewers who have helped improve the quality of this paper. This work was supported by the National Natural Science Foundation of China (Grant No. 61471028, 61571010, and 61702062), and the Fundamental Research Funds for the Central Universities (Grant No. 2017JBM004).

REFERENCES

[1] Administration, United States Federal Highway, Traffic monitoring guide, Evaluation, (2001).
[2] K. Aleksandra, K. Krishnaram, M. Nina and N. Alexandros, Releasing search queries and clicks privately, International Conference on World Wide Web, (2009), 171–180.
[3] J. Ankur, Adaptive stream resource management using Kalman Filters, ACM SIGMOD International Conference on Management of Data, (2004), 11–22.
[4] A. Blum, K. Ligett and A. Roth, A learning theory approach to non-interactive database privacy, Journal of the ACM, 60 (2008), Art. 12, 25 pp.
[5] C. A. Bradley, H. Rolka, D. Walker and J. Loonsk, BioSense: Implementation of a national early event detection and situational awareness system, Mmwr Supplements, 54 (2005), 11pp.
[6] Z. Cai and Z. He and X. Guan and Y. Li, Collective data-sanitization for preventing sensitive information inference attacks in social networks, IEEE Transactions on Dependable and Secure Computing, PP (2016), 1–1.
[7] J. Cao, Q. Xiao, G. Ghinita, N. Li, E. Bertino and T. Kian Lee, Efficient and accurate strategies for differentially-private sliding window queries, International Conference on Extending Database Technology, (2013), 191–202.
[8] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai and X. Li, Publishing setvalued data via differential privacy, Vldb, 4 (2012), 1087–1098.
[9] S. Cheng and Z. Cai and J. Li and H. Gao, Extracting kernel dataset from big sensory data in wireless sensor networks, IEEE Transactions on Knowledge and Data Engineering, 29 (2017), 813–827.
[10] S. Cheng and Z. Cai and J. Li, Curve query processing in wireless sensor networks, IEEE Transactions on Vehicular Technology, 64 (2015), 5198–5209.
[11] S. Cheng and Z. Cai and J. Li and X. Fang, Drawing dominant dataset from big sensory data in wireless sensor networks, The 34th Annual IEEE International Conference on Computer Communications (INFOCOM 2015), (2015), 531–539.
[12] D. Cynthia, Differential privacy, International Colloquium on Automata, Languages, and Programming, 4052 (2006), 1–12.
[13] D. Cynthia, Differential privacy in new settings, Acm-Siam Symposium on Discrete Algorithms, (2010), 174–183.
[14] C. Dwork, M. Naor and T. Pitassiand and G. N. Rothblum, Differential privacy under continual observation, STOC ’10 Proceedings of the Forty-Second ACM Symposium on Theory of Computing, (2010), 715–724.
[15] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum and S. Vadhan, On the complexity of differentially private data release: Efficient algorithms and hardness results, Proc. 41st Annu. ACM STOC, (2009), 381–390.
[16] C. Dwork, F. McSherry, K. Nissim and A. Smith, Calibrating noise to sensitivity in private data analysis, Theory of Cryptography, 3876 (2006), 265–284.
[17] L. Fan and L. Xiong, An adaptive approach to real-time aggregate monitoring with differential privacy, IEEE Transactions on Knowledge and Data Engineering, 26 (2014), 2094–2106.
[18] T. Florian, Privacy issues in vehicular ad hoc networks, International Conference on Privacy Enhancing Technologies, (2005), 197–209.
[19] B. J. Frey and D. Dueck, Clustering by passing messages between data points, Science, 315 (2007), 972–976.
[20] C. Graham, P. Cecilia, S. Divesh and T. T. L. Tran, Differentially private summaries for sparse data, International Conference on Database Theory, (2011), 299–311.
[21] Z. He and Z. Cai and J. Yu, Latent-data Privacy Preserving with Customized Data Utility for Social Network Data, IEEE Transactions on Vehicular Technology, 67 (2018), 665–673.
[22] Z. He, Z. Cai, S. Cheng and X. Wang, Approximate aggregation for tracking quantiles and range countings in wireless sensor networks, *Theoretical Computer Science*, 607 (2015), 381–390.

[23] Th. H. Chan, E. Shi and D. Song, Private and continual release of statistics, *Automata, Languages and Programming*, Part II, 405–417, Lecture Notes in Comput. Sci., 6199, Springer, Berlin, 2010.

[24] T. H. Hubert Chan, M. Li, E. Shi and W. Xu, Differentially private continual monitoring of heavy hitters from distributed streams, *Springer Berlin Heidelberg*, (2012), 140–159.

[25] R. Kalman, A new approach to linear filtering and prediction problem, *Journal of Basic Engineering*, 82 (1960), 35–45.

[26] G. Kellaris, S. Papadopoulos, X. Xiao and D. Papadias, Differentially private event sequences over infinite streams, *Proc. of the VLDB Endowment*, 7 (2014), 1155–1166.

[27] D. Kifer and B.-R. Lin, Towards an axiomatization of statistical privacy and utility, *Proc. of ACM PODS*, (2010), 147–158.

[28] C. Li, M. Hay, V. Rastogi, G. Miklau and A. Meggendorfer, Optimizing linear counting queries under differential privacy, *Twenty-Ninth ACM Sigmod-Sigact-Sigart Symposium on Principles of Database Systems*, (2010), 123–134.

[29] J. Li, S. Cheng, Z. Cai, J. Yu, C. Wang and Y. Li, Approximate holistic aggregation in wireless sensor networks, *2015 IEEE 35th International Conference on Distributed Computing Systems*, (2015).

[30] Y. Liang and Z. Cai and Q. Han and Y. Li, Location privacy leakage through sensory data, *Security and Communication Networks*, (2017), Article ID 7576307, 12 pages.

[31] J. Mao, W. Tian, Y. Zhang, J. Cui, H. Ma, J. Bian, J. Liu and J. Zhang, Co-Check: Collaborative outsourced data auditing in multicloud environment, *Security and Communication Networks*, (2017), Article ID 2948025, 13 pages.

[32] J. Mao, Y. Zhang, P. Li, T. Li, Q. Wu and J. Liu, A Position-aware Merkle Tree for Dynamic Cloud Data Integrity Verification, *Soft Computing*, 21 (2017), 2151–2164.

[33] H. Michael, R. Vibhor, M. Gerome and D. Suciu, Boosting the accuracy of differentially private histograms through consistency, *Proceedings of the VLDB Endowment*, 3 (2010), 1021–1032.

[34] J. Sun and Y. Fang and X. Zhu, Privacy and emergency response in e-healthcare leveraging wireless body sensor networks, *IEEE Wireless Communications*, 17 (2010), 66–73.

[35] B. Thomas, A framework for generating network-based moving objects, *GeoInformatica*, 6 (2002), 153–180.

[36] Q. Wang, Y. Zhang, X. Lu and Z. Wang, RescueDP: Real-time spatio-temporal crowdsourced data publishing with differential privacy, *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, (2016).

[37] X. Xiao, G. Bender, H. Michael and G. Johannes, iReduct: differential privacy with reduced relative errors, *ACM SIGMOD International Conference on Management of Data, SIGMOD 2011, Athens, Greece, June*, (2011), 229–240.

[38] J. Xu, Z. Zhang, X. Xiao, Y. Yang and G. Yu, Differentially private histogram publication, *IEEE International Conference on Data Engineering*, (2012), 32–43.

[39] L. Zhang and Z. Cai and X. Wang, FakeMask: A Novel Privacy Preserving Approach for Smartphones, *IEEE Transactions on Network and Service Management*, 13 (2016), 335–348.

[40] X. Zheng and Z. Cai and J. Yu and C. Wang and Y. Li, Follow but no track: Privacy preserved profile publishing in cyber-physical social systems, *IEEE Internet of Things Journal*, 4 (2017), 1868–1878.

[41] CTR Data. Available from: https://www.kaggle.com/c/avazu-ctr-prediction.

[42] Capital Bikeshare Data. Available from: https://www.capitalbikeshare.com/system-data.

Received November 2017; revised February 2018.

E-mail address: 16120158@bjtu.edu.cn
E-mail address: yhuo@bjtu.edu.cn
E-mail address: hcq0394@gmail.com
E-mail address: yflu@bjtu.edu.cn
E-mail address: 14213008@bjtu.edu.cn