Local Differential Privacy for Data Streams

Xianjin Fang, Qingkui Zeng, and Gaoming Yang[✉]

School of Computer Science and Engineering,
Anhui University of Science and Technology, Huainan, China
gmyang@aust.edu.cn

Abstract. The dynamic change, huge data size, and complex structure of the data stream have made it very difficult to be analyzed and protected in real-time. Traditional privacy protection models such as differential privacy which need to rely on the trusted servers or companies, and this will increase the uncertainty of protecting streaming privacy. In this paper, we propose a new privacy protection protocol for data streams under local differential privacy and w-event privacy, which makes it possible to keep up-to-date statistics over time, and it is still available when the third parties are untrusted. We use sliding window to collect the data streams in real-time, finding out the occurrence of significant moves, capturing the latest data distribution trend, and releasing the perturbed data streams report in time. This protocol provides a provable privacy guarantee, reduces computation and storage costs, and provides valuable statistical information. The experimental results of real datasets show that the proposed method can protect the privacy of the data streams and provide available statistical data at the same time.

Keywords: Data streams · Local differential privacy · w-event privacy · Sliding window

1 Introduction

With the development of 5G technology, intelligent devices and sensors have produced more and more dynamic data, which we call the data stream. Real-time analysis of stream data can obtain valuable information to understand an important phenomenon [13], so it is widely used in various application fields, such as mobile crowd sensing [28], traffic service stream monitoring [19] and social network hotspot tracking [26]. The data service providers collect real-time data stream and publish real-time statistics, share and analyze [29] them with interested third-party to improve the service quality.

This work was supported by National Natural Science Foundation of China (61572034), Major Science and Technology Projects in Anhui Province (18030901025), Anhui Province University Natural Science Fund (KJ2019A0109).

S. Yu et al. (Eds.): SPDE 2020, CCIS 1268, pp. 143–160, 2020.
https://doi.org/10.1007/978-981-15-9129-7_11

However, there are potential privacy risks in this process. On account of the joining of the untrusted third party, the attacker may query the original data of multiple timestamps of a single user through the differential attack to draw the user's data track and disclose the user's privacy information [20]. Recently research [4] has found that the user's mobile trajectory is highly unique from the user's mobile data obtained by mobile phone operators. Even if the desensitized dataset provides a small amount of anonymous information, it can still be linked to the designated user with relevant background knowledge. A series of similar findings reveal that the privacy of personal data stream is facing a huge risk, so it is of great significance to the research and development of data stream privacy collection and release mechanism, but in the real-time, irreversibility and large scale of data stream itself also bring challenges to the research.

Differential privacy (DP) [10] as a widely used privacy protection model provides strict privacy guarantee and theoretical proof, and it does not need to consider the attacker's background knowledge. One of the common methods of data publishing with differential privacy is to perturb the data before publishing and hide the sensitive information of individuals in the process of statistical analysis and data mining. At present, the research on the differential privacy model mainly focuses on the static scene, however the real-time data is collected and published all in the dynamic scene [5]. In the centralized interactive scene of differential privacy, the trusted curators collect sensitive data from different entities, carefully adding calibrated noise, and then sharing the final results with data analysts. The model is shown in Fig. 1. Dwork et al. [11] proposed two different privacy schemes for continuous data collection, namely event-level and user-level privacy. Event-level privacy protects user's privacy on a single timestamp in the data stream, but it does not protect user's privacy in the whole data stream; user-level privacy needs to add noise in the whole data stream, which will reduce the utility of data in the long run.

This model requires that the trusted data curator, however, if the curator is not trusted, there is a risk of the potential breach of privacy from a third party. And the attacker may obtain part of the original data of the data curator through repeated queries and infer the user's privacy. Local Differential Privacy (LDP) [21] is a distributed variant of differential privacy, which does not require a trusted data curator. Before sending individual privacy data, users perturb their data on the local device and send the perturbed privacy reports to resist privacy attacks under the centralized model. This model is shown in Fig. 2. At now, the research of the local differential privacy model is mainly focused on the release of single data [3], however, it's difficult for the LDP model to deal with complex real-time data. As the streaming evolves, the consumption of computing power and storage space will become larger, and the privacy budget will gradually decline.

Compared with the collection of traditional data, the length and content of the data streams change dynamically, the data size is huge, and the data type is complex. Aiming at the problem of how to protect streaming privacy under the local differential privacy model, we propose the locally differential

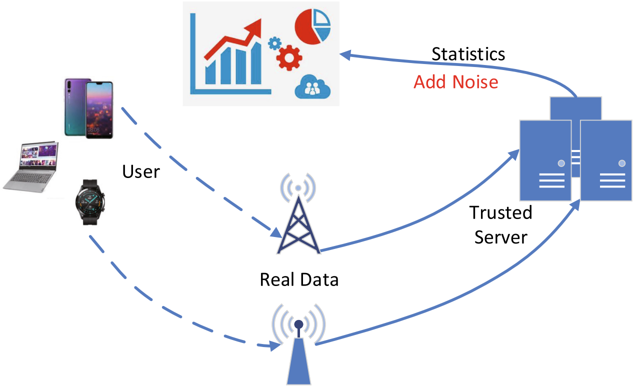


Fig. 1. Differential privacy protection model.

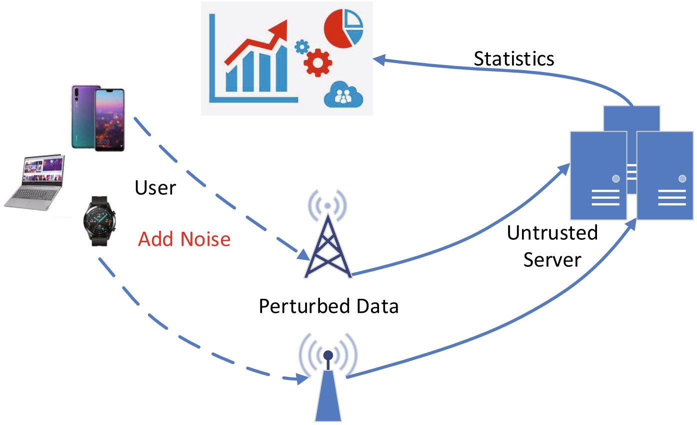


Fig. 2. Local differential privacy protection model.

private Streaming (LDPS) protocol based on local differential privacy and w-event privacy [22]. The LDPS protocol selects different algorithms to obtain the statistics according to the type of data streams. When the collected data streams are real-time numerical attributes such as temperature and humidity, longitude, latitude, and heart rate, etc., we calculate the value of mean [25]; for real-time classified attributes such as the user's default browser home page, search engine setting and most frequently emojis or words, etc. [27], we conduct the frequency estimation and find heavy hitters [2]. When data types are mixed, different protocols are used to process the data with different attributes identified. The main contributions are:

(1) Ensure that the individual user data never leaves the device by deploying the local differential privacy;

(2) The proposed protocol provides a stronger privacy guarantee for the data stream and reduces the ways for attackers to breach privacy, meanwhile, obtains valuable statistical data;
(3) The sliding window technology is used to release the private streaming in real-time, which reduces the computing and storage overhead and privacy budget consumption compared with the traditional privacy protection method;
(4) To further reduce the storage space and better allocate the privacy budget, the proposed protocol determine the window length of the stable sub-stream adaptively, detect the occurrences of significant moves and open a new window in time to capture the trend and distribution of data streams.

This paper is organized as follows. First, we describe related works in Sect. 2 and the background knowledge for this paper in Sect. 3. Then, we state the problem setting and methodology in Sect. 4 and propose our protocol in Sect. 5. Next, we evaluate and analyze our method in Sect. 6. At last, the work is summarized in Sect. 7.

2 Related Works

Recently, the research schemes for the privacy protection of data streams are mainly focused on the release of real-time time series under different privacy budgets. Fan et al. [15] propose a framework of FAST based on perturbation, filtering, and sampling. According to the error rate between the estimated and predicted statistical data, the framework releases noisy report at the sampling points, which can provide user-level privacy protection, i.e., to protect the privacy of the user in the whole time-series. However, their work cannot be applied to infinite data streams because the FAST must allocate the maximum number of releases in advance, and the sampling mechanism can only be applied if each timestamp has an equivalent budget. Kellaris et al. [22] propose a new model, w-event ϵ-differential privacy (w-event privacy for short), which combines the gap between event-level privacy and user-level privacy, they also give new mechanisms to implement the w-event privacy model.

Differential privacy has attracted much attention in the real-time release of streaming data [17] because of its advantages in mathematical proof and privacy protection. However, these mechanisms are based on trusted servers, which strictly limits their application in practice. Fan et al. [14] propose an adaptive system, which releases aggregate statistical information of real-time and spatio-temporal data streams under differential privacy model by sampling and filtering steps. Although this mechanism optimizes the budget allocation of numerical attributes, it applies only to finite data streams. Wang et al. [30] present the adaptive framework AdaPub, which can update the parameters with the data stream evolving. These researches extend the mechanism by considering the sliding window of the w timestamp and optimizing the budget allocation within the window. While these efforts provide good insight into publishing data streams under differential private guarantees, they rely on a trusted server that is not convenient to deploy in many real-world applications.

In order to solve the problem of untrusted servers, many scholars and researchers discuss the local differential privacy model, i.e., the individual raw data is perturbed before it is sent from the client. Duchi et al. [6–8] proposed the min-max mechanism of numerical attribute publication based on local differential privacy. Erlingsson et al. [12] developed a RAPPOR protocol for real-time publishing binary attributes, which is based on random response technology to limit the probability of inference of sensitive information. Wang et al. [12] improve the accuracy of numerical attributes of the min-max mechanism and extend it to publish binary and numerical attribute data. However, the proposed mechanism randomly selects k attributes for perturbation, which is not realistic in some practical application scenarios. Kim et al. [23] develop a mechanism for the health data stream by leveraging local differential privacy. In addition, these mechanisms cannot be used to distinguish between ordered and disordered attributes.

The mechanisms mentioned above carefully allocate the privacy budget on each timestamp. However, even in a relatively short period, repeated differential privacy computing will accumulate the privacy loss to a large value, so an adaptive compression mechanism is needed to reduce the loss of privacy budget. Recently, Joseph et al. [20] apply a compression technique to continuously release binary attributes under local differential privacy. For user clients with similar data distribution, this mechanism will consume the local privacy budget only when the distribution of users changes significantly. Soheila et al. [13] propose an adaptive dynamic compression method in the local differential privacy data stream mechanism, which adaptively adjusts the window length to reduce the consumption of the privacy budget. Wang et al. proposed a RescueDP protocol [29], which provides privacy protection statistical data distribution on infinite timestamps through adaptive sampling, adaptive budget allocation, dynamic grouping, perturbation, and filtering mechanisms.

3 Preliminaries

In this section, we introduce the local differential privacy and w-event privacy model, and some related techniques used in these models. The definitions used in this paper are given below.

Definition 1 *(w-neighboring). Let $S_t = \{D_1, D_2, \ldots, D_t\}$ be a prefix stream of sequential data where at each timestamp i, a dataset D_i is collected with an arbitrary number of rows each corresponding to a unique user. For any positive integer w, two prefix streams S_t, S_t' are defined as w-neighboring if:*

(1) for each D_i, D_i', $i \in [1, t]$ and $D_i \neq D_i$' it holds that D_i , D_i' neighboring, and;
(2) for each $i_1 \in [1, t]$, $i_2 \in [1, t]$, $i_1 < i_2$, and $i_1 \neq i_2$, it holds that $i_2 - i_1 + 1 \leq w$.

The sliding window arranges tuples in streaming data according to their timestamps. The sliding window, with a fixed length w, always keeps the newest

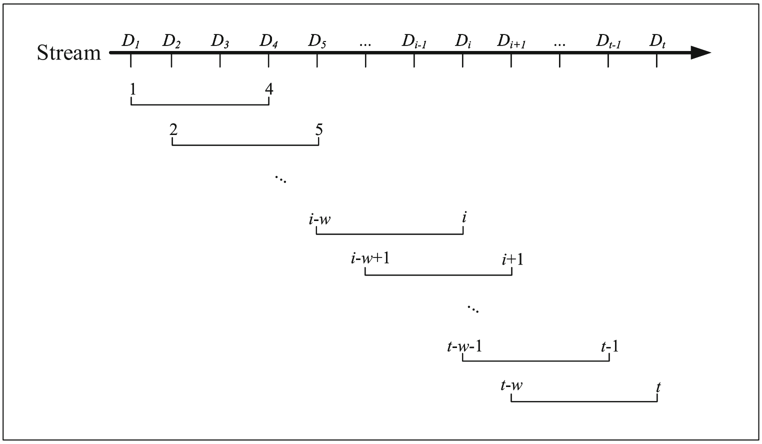

Stream
D_1 D_2 D_3 D_4 D_5 ... D_{i-1} D_i D_{i+1} ... D_{t-1} D_t
1
4
2
5
i-w
i
i-w+1
i+1
t-w-1
t-1
t-w
t

Fig. 3. Sliding window model.

$w - th$ tuples, while discarding old ones. Figure 3 shows the sliding window model with the window length of 4. w-event privacy is an extension of differential privacy for continuously publishing data streams. It provides a provable privacy guarantee for any sequence of events that occur in any sliding window of the w timestamp. The definition as follows:

Definition 2 *(w-event privacy). Let M be a random mechanism, and let D be the domain of all possible output M. M satisfies w-event ϵ-differential privacy when S_t and S_t' are w-neighboring, if it holds that:*

$$Pr[M(S_t) \in D] \leq e^{\epsilon} \cdot Pr[M(S_t') \in D] \tag{1}$$

Definition 3 *(Stably sub-stream). Given a threshold δ >0, a sub-stream is stably from timestamp i to j, if and only if d(S_t, S_i') $\leq \delta$, $\forall$t, t'$\in$[i, j], where d($\cdot$) is a distance measurement method.*

The generation of a significant move denotes the outbreak of a new event or the occurrence of a new trend in the data stream. When the newly observed tuples maintain the stability of the current window, they can be added to the current window to form a stably sub-stream. Otherwise, a new window should be opened to capture new trends in the data stream.

Definition 4 *(Significant move). Let $S_{i,j}$ be a stably sub-stream in timestamp [i, j]. A newly observed tuple S_{j+1} is a significant move if the distance of S_j and S_{j+1} is greater than δ.*

Local differential privacy is a new privacy definition for individual privacy in clients and a special case of differential privacy, the perturbation process in LDP shifts from the server-side to the client-side. Under this definition, the modification between any two pieces of local data has little impact on the query results.

Even if a piece of data is known, an adversary cannot obtain accurate individual information by observing the query results because of data perturbations on the local client. Therefore, the risk of privacy disclosure between the two local data is in a very small and acceptable range. The definition as follows:

Definition 5 *(ϵ-local differential privacy). Where ϵ>0, a randomized mechanism A satisfies ϵ-local differential privacy if and only if, for any pairs of input tuples x and x', for any possible output x^* (in the domain belonging to A), we have:*

$$Pr[A(x) = x^*] \leq e^{\epsilon} \cdot Pr[A(x') = x^*] \tag{2}$$

Theorem 1 *(Sequential composition). Consider mechanism A that provides ϵ_i-local differential privacy. A sequence of mechanism A over a data stream S provides $\sum \epsilon_i$-local differential privacy.*

4 Problem Setting

In this section, we discuss the problem statement about the data stream under the local differential privacy model and propose some methods to solve these problems.

4.1 Problem Statement

We describe the data stream firstly. We consider an infinite source stream dataset S of d states and denote the stream that collected from the user set U_i in first timestamp $_i$ as S_i, $S_i = \{D_1, D_2, ..., D_i\}$, $i \in [1, t]$, D_i is the dataset sent by U_i's users at timestamp i. We set the data stream within the timestamp range i to j as $S_{i,j}$.

To protect the privacy of real-time data streams under the limited storage space and computing power of edge nodes, the client's data streams must be perturbed before being sent to the server to ensure privacy requirements. Therefore, our goal is to publish the infinite data streams in real-time, which can ensure the privacy of each client, maintain the data utility and provide valuable statistical information.

First, We need to prevent privacy breaches before the client transfers the data stream. Second, how to use the local differential privacy model to ensure the utility of data while providing privacy protection for the individual data and how to reduce the computing power and large overhead of storage space caused by the perturbed mechanism. Third, how to detect the concept drift of data streams to reduce errors. Last, how to meet the definition of w-event privacy and adjust the privacy budget allocation adaptively in real-time under the local differential privacy model.

4.2 Propose Solution

The individual raw data stream will never leave the devices through the deployment of the local differential privacy model. This will provide a powerful privacy guarantee and reduce how adversaries can breach privacy.

Our goal is to protect raw individual data and to publish the perturbed stream which satisfies w-event ϵ-differential privacy while providing valuable statistics. With the evolving of the data stream, the stable subseries and judges the significant points appear, we will sniff out new trends of the data streams in time and allocates the privacy budget adaptively. Therefore, this method can be applied to infinite data streams and reduce storage space and computing consumption.

The privacy protection of numerical attribute data streams consists of four steps: standardization, perturbation, adaptive allocation, and decoding. The first step is to standardize the numerical attribute data stream and encode the normalized data according to the corresponding mechanism. The second part is the perturbation, which implements the perturbed mechanism of the standardized data stream that satisfies the definition of local differential privacy. The third step is adaptive allocation, which determines the stably sub-stream and distinguishes the significant moves in time, and dynamically adaptively allocates the privacy budget. In the fourth step, data streams after the real-time perturbation are collected, and the value of mean is obtained after aggregation, and the mean value is normalized and restored.

The privacy protection process of categorical attribute data streams consists of four steps: encoding, perturbation, adaptive allocation, and decoding. The first step is to encode the data stream, such as a one-hot encoding or bloom filters. The second step is to deal with the perturbation of the encoded data stream which satisfies the definition of local differential privacy. The third step is to compare the data in the sliding window during the perturbation and allocate the privacy budget adaptively according to the data distribution of the sliding window to make it meet the definition of w-event privacy. The fourth step is to decode the aggregated data stream and get the frequency estimation of the classification attributes.

5 The Local Differentially Private Streaming Protocol

In this section, we describe our mechanism Local Differentially Private Streaming (LDPS) for publishing multi-variable data streams under local and w-event differential privacy.

5.1 Numerical Attributes

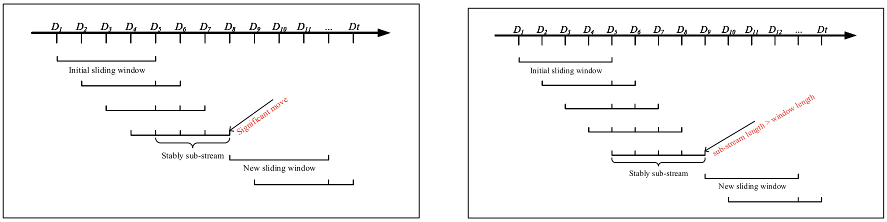
For numerical attributes, our goal is to estimate the mean value from the sanitized stream. We standardize the raw streams S, set the sliding window with the length of w, threshold t and privacy budget ϵ. The perturbation mechanism

Algorithm 1. The Local Differentially Private Streaming for numerical attributes

```
Input: the streams $S$; the window length $w$; the threshold $t$; and privacy budget $\epsilon_i$;
Output: $S\_mean$
$Normalize\ S$;
Initial $Per = [], count = 0$;
for $i = 0$ to $w$ do
    $Per.append(LDP(S_i))$
end for
for $i = w$ to $n$ do
    if $count = w)$ then
        $count = 0$
        $Per.append(LDP(S_i))$
        continue
    end if
    if $count < w)$ then
        if $d(d(S_{i-2} - S_{i-1}) - d(S_{i-1} - S_i)) \leq t$ then
            $Per.append(Per[i-1])$
            $count+ = 1$
        else
            $Per.append(LDP(S_i))$
            $count = 0$
        end if
    end if
end for
$S\_mean = Decode(Per)$;
$Denormalize(S\_mean)$;
return $S\_mean$;
```

of the proposed method follows the typical LDP numerical attribute mechanism and parameters.

The LDP perturbation mechanism is conducted normally in initial w timestamps. After timestamp w, the sliding window will slide and allocate the privacy budget adaptively with the stream evolving. We calculate the l_1 distance between the real value of the current timestamp and the real value of the previous and subsequent timestamps respectively and set the threshold t to compare the difference between these two distances. If the difference is less than the threshold, then the perturbed report at the subsequent timestamp will be the same one in the current timestamp, and the stably sub-stream is formed; if the difference is greater than the threshold, we will continue the perturbation mechanism at the subsequent timestamp and denote this timestamp as a significant move. When a significant move occurs as shown in Fig. 4(a) or the length of stably sub-stream is greater than the sliding window length w as shown in Fig. 4(b), a new window is opened and the LDP perturbation steps are conducted for the next w timestamps.



(a) The model when a significant point occurs.

(b) The model when the sub-stream length is greater than the window length.

Fig. 4. The Local differentially private streaming protocol for numerical attributes.

The server collects the perturbed report of sequence S and aggregates it, calculates all the perturb report mean value and reverses standardization to get the estimated mean value. Specific steps such as Algorithms 1.

5.2 Categorical Attributes

For categorical attributes, our goal is to estimate each value frequency from the sanitized stream. Given the input streams S and the sliding window length w, the perturbation mechanism of the proposed method follows the chosen LDP categorical attribute mechanism and parameters.

We conduct the LDP perturbed mechanism in initial w timestamps. In the next timestamp, the sliding window will begin to slide with the time-series S. If the real data of the subsequent timestamp has been released in its previous w timestamps sliding window, then the perturbed report will be the same one in the current timestamp. If the real data of the subsequent timestamp has not been released in the previous w timestamp, we will continue the perturbation mechanism at the subsequent timestamp and denote this timestamp as a significant move.

The server collects the perturbed report of sequence S and aggregates it. The sub-stream of perturbed data in the sliding window is w nearest neighbor data stream, which satisfies the definition of w-event privacy, conforms to the differential privacy combination theorem, and satisfies the definition of LDP between each two tuples. Specific steps such as Algorithms 2.

5.3 Privacy Analysis

We first prove the LDPS protocol satisfies the ϵ-local differential privacy, and then prove that it also satisfies the w-event privacy.

Let S_i be the current timestamp data stream and S_{i+1} be the last time release. To prove that the LDPS protocol satisfies w-event privacy, first, we need to prove that the perturbed report for every two timestamps satisfies the definition of ϵ_i-local differential privacy. Then, according to Theorem 1, we need to prove that the sum of the privacy budgets consumed by the LDPS within a window of length w does not exceed.

Algorithm 2. The Local Differentially Private Streaming for categorical attributes

```
Input: the streams $S$; the window length $w$; the perturbed mechanism parameters;
Output: $f(d_i), i \in (0, |D|)$
InitialPer[];
for $i = 0$ to $w$ do
    $Per.append(LDP(S_i))$
end for
for $i = w$ to $n$ do
    for $j = i - w - 1; j > i - 1; j - -$ do
        if then$S_j = S_i$
            $Per.append(LDP(S_j))$
            Break
        else
            $Per.append(LDP(S_i))$
        end if
    end for
end for
$f(d_i) = Decode(Per)$;
return $S\_mean$;
```

Theorem 2. *The Local Differentially Private Streaming protocol satisfies ϵ-local differential privacy.*

Proof. In the perturbation step of LDPS, we perturb the streaming on clients by the LDP's mechanism. So, we can ensure that the tuples in the neighboring timestamp are satisfied the ϵ-local differential privacy.

Theorem 3. *The Local Differentially Private Streaming protocol satisfies w-event privacy.*

Proof. Due to the allocation of privacy budget adaptively, we compress the privacy budget to 0 which using the same perturbation report as at the previous timestamp. Assume that the i-th data privacy budget is ϵ_i, we consider the definition of local differential privacy protection in two scenarios respectively. First, in the scenario that perturbation at both the current and previous timestamp, we use the same LDP perturb mechanism with the same parameters and therefore have the same privacy budget. In the timestamp i, by definition of ϵ-local differential privacy we have:

$$\frac{Pr[LDPS(S_i) = x^*]}{Pr[LDPS(S_{i-1}) = x^*]} \leq e^{\epsilon_i} \tag{3}$$

Second, in the scenario that the current timestamp stream adopts the previous perturb report, the current privacy budget is adaptively compressed to 0, and the privacy guarantee begins to decline because of releasing repeatedly, we have:

$$\frac{Pr[LDPS(S_i) = x^*]}{Pr[LDPS(S_{i-1}) = x^*]} = 1 \leq e^0 \tag{4}$$

So, the LDPS protocol satisfies the ϵ_i-local differential privacy in differential scenarios.

In the whole sliding window, there are also two scenarios: normal perturbation in the whole window and form stably sub-stream in the window. We set each tuple in the whole sliding window has itself privacy budget ϵ_i, even if the budget is 0. According to theorem 1, the data stream in the current sliding window satisfies $\sum \epsilon_i$-local differential privacy. So, we have a privacy budget of $\sum \epsilon_i$ for w-event $\sum \epsilon_i$ differential privacy, by definition of w-event privacy, there are:

$$\frac{Pr[LDPS(S_t) \in D]}{Pr[LDPS(S_i') \in D]} \leq e^{\sum \epsilon_i} \tag{5}$$

6 Experiments

6.1 Experimental Setup

Datasets. We selected three public datasets as experimental datasets.

Table 1. Experimental datasets.

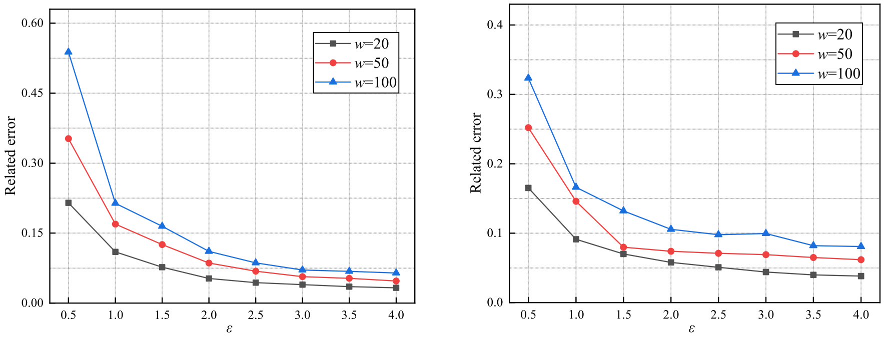
Dataset	IPUMS	Twitter daily activities	Gas sensor
Number of Instances	1000000	60093175	919438
Domain size/Mean value	78	635	27.1767, 57.5680

We choose the 2017 *Integrated Public Use Microdata Series* (IPUMS) [1] and selects the age attribute, which has 25 data categories; we extract 1% from the dataset and take the first million pieces of data as the experimental dataset for the categorical attribute. *Twitter daily activities* [24] is the Microsoft Research datasets of longitudinal, daily, per-county activity periods of aggregated Twitter users. We extract 500000 records and choose the per-country attribute as experimental datasets for the categorical attribute.

The *Gas-Sensor* dataset [18] has recordings of a gas sensor array composed of 8 MOX gas sensors, and a temperature and humidity sensor. We use the humidity and temperature attribute as the experimental datasets for the numerical attribute. The number of instances, domain sizes, or mean values of datasets are shown in Table 1.

Experimental Situation. These experiments were implemented in Python 3.7 with NumPy and xxhash libraries and were performed on a PC with Intel Core i7-7700hq CPU and 16 GB RAM. Each experiment was repeated 100 times to reduce the influence of contingency on the experimental results.

Parameter Setting. We consider varying the privacy budget parameter ϵ, the length of sliding window w and threshold t for mean value computing and varying the length of sliding window w for frequency estimation. In the mean estimation experiment, Duchi et al.'s method [9] and the Laplace mechanism [31] are chosen as the LDP perturbation mechanism. For convenience, respectively, they are abbreviated as Duchi and LM. In the frequency estimation experiment, RAPPOR [16] mechanism is chosen as the LDP perturbation mechanism. We adopt the same parameter settings and noise correction methods when using these typical LDP mechanisms.



(a) The effects of privacy budgets in LM.

(b) The effects of privacy budgets in Duchi's method.

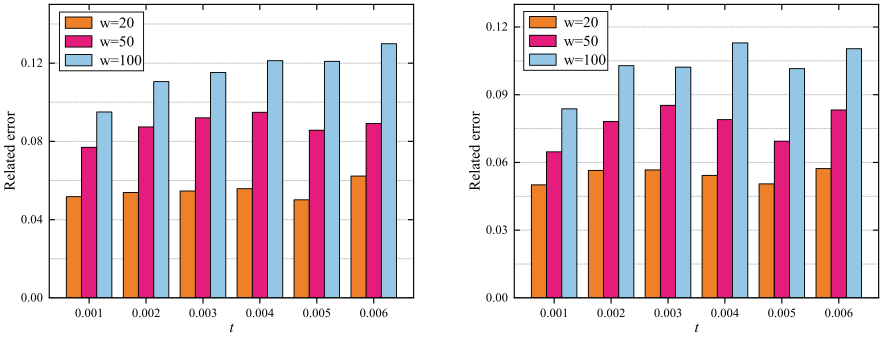
Fig. 5. The effects of privacy budgets on temperature.

Experimental Metrics. Related error is taken as the error measure of mean value calculation, and MAPE is taken as the error measure of frequency estimation. The related error is the absolute value of the predicted value minus the real value divided by the real value. The definition of related error is as follows:

$$Realtederror = \left|\frac{y_i - x_i}{x_i}\right| \times 100\% \tag{6}$$

And the MAPE is the absolute value between the estimated and true frequency, divide the absolute value by the true frequency, then cumulate these values and divide by the size of the data value domain. The definition of MAPE is as follows:

$$MAPE = \frac{\sum_{i=1}^{|D|}\left|\frac{y_i - x_i}{x_i}\right|}{|D|} \times 100\% \tag{7}$$



(a) The effects of thresholds in LM.

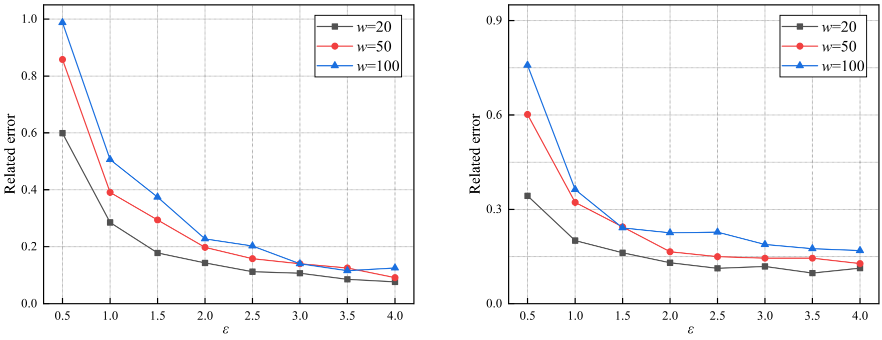
(b) The effects of thresholds in Duchi's method.

Fig. 6. The effects of thresholds on temperature.

6.2 Results for Mean Value

To evaluate the data utility of LDPS on numerical attributes, we calculate the mean value of the temperature and humidity attributes of the Gas-sensor dataset respectively by varying ϵ, threshold t, length of window w. Related error in Eq. (6) is selected as a metric.

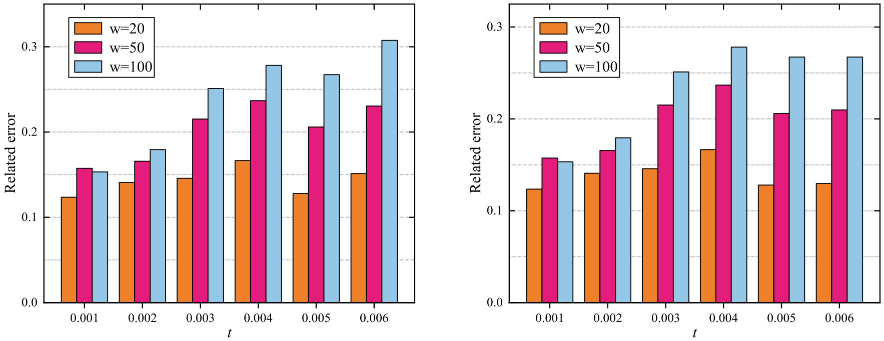
We choose w to be 20, 50, 100, ϵ from 0.5 to 4.0, and t from 0.001 to 0.006. When we evaluate the effects of different ϵ values, make t 0.003; evaluate the effects of different t, make ϵ 2. Figure 5(a) shows the effects of ϵ and w on data utility when the Duchi's method is chosen as perturbing mechanism, and Fig. 5(B) shows the effects of ϵ and w on data utility when the Laplace mechanism is chosen as perturb mechanism. It can be seen that the data utility is higher



(a) The effects of thresholds in LM.

(b) The effects of privacy budgets in Duchi's method.

Fig. 7. The effects of privacy budgets on humidity.



(a) The effects of thresholds in LM.

(b) The effects of thresholds in Duchi's method.

Fig. 8. The effects of thresholds on humidity.

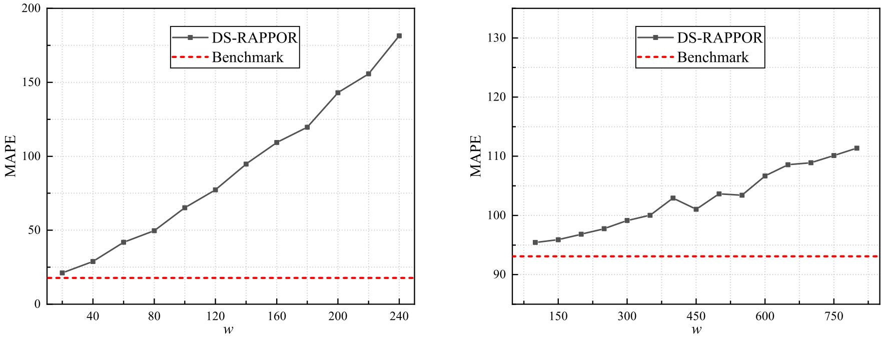
when ϵ is large and w is small; the data utility is lower when ϵ is small and w is large, while the privacy guarantee level is on the contrary.

In Fig. 6, we can see the effects of t and w under the Duchi and LM perturb mechanism. As t value increases, data utility does not always decline, but increases first and then decreases. And the effects of window length is similar to the results in Fig. 5, The higher the w, the lower the data utility.

We verify the data utility of the same parameters on the humidity attribute dataset. The results are shown in Fig. 7 and Fig. 8, and we can get similar effects on data utility.

6.3 Results for Frequency Estimation

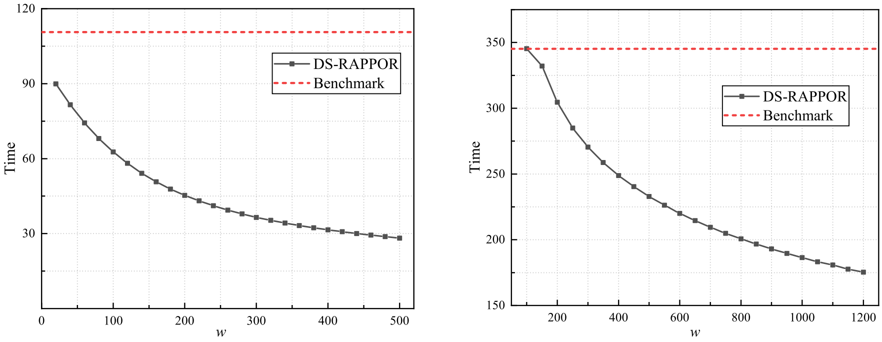
To evaluate the data utility of LDPS on categorical attributes, we estimate the frequency of each attribute value in the 'AGE' attribute of IPUMS and



(a) The effects of w on 'AGE'.

(b) The effects of w on 'CountryID'.

Fig. 9. The effects of w in frequency estimation.



(a) The effects of w on 'AGE'.

(b) The effects of w on 'CountryID'.

Fig. 10. The effects of runtime in frequency estimation.

'CountryID' attribute of Twitter daily activities by varying the length of window w and choose MAPE in Eq. (7) as a metric. The parameters k, h, p, q and f of the perturbation mechanism RAPPOR are set to 256, 4, 0.5, 0.75, 0.5 on 'AGE' and 256, 8, 0.5, 0.75, 0.5 on 'CountryID', respectively.

We consider conducting experiments from the perspectives of data utility and runtime and choose RAPPOR protocol as the perturb mechanism. And we change the time-series streaming to normal data, adopt the RAPPOR to perturb these data as a benchmark. Figure 9 shows that the MAPE value increases with the increase of the window length w. On the contrary, Fig. 10 shows that the experimental runtime decreases with the increase of the window length w. So, we can see that varying the window length w affect data utility, privacy, and runtime.

7 Conclusion

This paper focuses on the privacy protection of data streams. The untrusted third parties may query the original data of multiple timestamps of a single user to breach the user's privacy while current local differential privacy protocols can hardly handle the data streams. We propose the local differentially private streaming protocol, which can not only protect streaming privacy but also ensure high utility, and less storage and computational power overhead. The proposed method utilizes the sliding window that satisfies w-event privacy to find the stably sub-stream and significant moves in real-time. The experimental results show that the proposed protocol has high utility, is suitable for both numerical and categorical attributes, and maintains its utility under different distributions and streaming sizes.

References

1. University of Minnesota: IPUMS USA. https://www.ipums.org
2. Bassily, R., Nissim, K., Stemmer, U., Thakurta, A.G.: Practical locally private heavy hitters. In: Advances in Neural Information Processing Systems, pp. 2288–2296 (2017)
3. Bassily, R., Smith, A.: Local, private, efficient protocols for succinct histograms. In: Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, pp. 127–135 (2015)
4. De Montjoye, Y.A., Hidalgo, C.A., Verleysen, M., Blondel, V.D.: Unique in the crowd: the privacy bounds of human mobility. Sci. Rep. **3**, 1376 (2013)
5. Ding, B., Kulkarni, J., Yekhanin, S.: Collecting telemetry data privately. In: Advances in Neural Information Processing Systems, pp. 3571–3580 (2017)
6. Duchi, J., Wainwright, M.J., Jordan, M.I.: Local privacy and minimax bounds: sharp rates for probability estimation. In: Advances in Neural Information Processing Systems, pp. 1529–1537 (2013)
7. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy and statistical minimax rates. In: 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, pp. 429–438. IEEE (2013)
8. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Local privacy, data processing inequalities, and statistical minimax rates. arXiv preprint arXiv:1302.3203 (2013)
9. Duchi, J.C., Jordan, M.I., Wainwright, M.J.: Minimax optimal procedures for locally private estimation. J. Am. Stat. Assoc. **113**(521), 182–201 (2018)
10. Dwork, Cynthia., McSherry, Frank., Nissim, Kobbi, Smith, Adam: Calibrating noise to sensitivity in private data analysis. In: Halevi, Shai, Rabin, Tal (eds.) TCC 2006. LNCS, vol. 3876, pp. 265–284. Springer, Heidelberg (2006). https://doi.org/10.1007/11681878_14
11. 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, pp. 715–724 (2010)
12. Erlingsson, Ú., Pihur, V., Korolova, A.: Rappor: randomized aggregatable privacy-preserving ordinal response. In: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, pp. 1054–1067 (2014)
13. Ezabadi, S.G., Jolfaei, A., Kulik, L., Kotagiri, R.: Differentially private streaming to untrusted edge servers in intelligent transportation system. In: 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), pp. 781–786. IEEE (2019)
14. Fan, L., Xiong, L.: An adaptive approach to real-time aggregate monitoring with differential privacy. IEEE Trans. Knowl. Data Eng. **26**(9), 2094–2106 (2013)
15. Fan, L., Xiong, L., Sunderam, V.: Fast: differentially private real-time aggregate monitor with filtering and adaptive sampling. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, pp. 1065–1068 (2013)
16. Fanti, G., Pihur, V., Erlingsson, Ú.: Building a rappor with the unknown: privacy-preserving learning of associations and data dictionaries. Proc. Priv. Enhancing Technol. **2016**(3), 41–61 (2016)
17. Hassidim, A., Kaplan, H., Mansour, Y., Matias, Y., Stemmer, U.: Adversarially robust streaming algorithms via differential privacy. arXiv preprint arXiv:2004.05975 (2020)

18. Huerta, R., Mosqueiro, T., Fonollosa, J., Rulkov, N.F., Rodriguez-Lujan, I.: Online decorrelation of humidity and temperature in chemical sensors for continuous monitoring. Chemom. Intell. Lab. Syst. **157**, 169–176 (2016)
19. Jolfaei, A., Kant, K.: Privacy and security of connected vehicles in intelligent transportation system. In: 2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks-Supplemental Volume (DSN-S), pp. 9–10. IEEE (2019)
20. Joseph, M., Roth, A., Ullman, J., Waggoner, B.: Local differential privacy for evolving data. In: Advances in Neural Information Processing Systems, pp. 2375–2384 (2018)
21. Kasiviswanathan, S.P., Lee, H.K., Nissim, K., Raskhodnikova, S., Smith, A.: What can we learn privately? SIAM J. Comput. **40**(3), 793–826 (2011)
22. Kellaris, G., Papadopoulos, S., Xiao, X., Papadias, D.: Differentially private event sequences over infinite streams (2014)
23. Kim, J.W., Jang, B., Yoo, H.: Privacy-preserving aggregation of personal health-data streams. PloS One **13**(11) (2018)
24. Microsoft: Longitudinal, daily, per-county activity periods of aggregated Twitter users. https://www.microsoft.com/en-us/download/details.aspx?id=57387
25. Nguyên, T.T., Xiao, X., Yang, Y., Hui, S.C., Shin, H., Shin, J.: Collecting and analyzing data from smart device users with local differential privacy. arXiv preprint arXiv:1606.05053 (2016)
26. Sun, L., Ge, C., Huang, X., Wu, Y., Gao, Y.: Differentially private real-time streaming data publication based on sliding window under exponential decay. Comput. Mater. Continua **58**(1), 61–78 (2019)
27. Team, A., et al.: Learning with privacy at scale. Apple Mach. Learn. J. **1**(8) (2017)
28. 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 Secure Comput. **15**(4), 591–606 (2016)
29. Wang, Q., Zhang, Y., Lu, X., Wang, Z., Qin, Z., Ren, K.: Rescuedp: real-time spatio-temporal crowd-sourced data publishing with differential privacy. In: IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications, pp. 1–9. IEEE (2016)
30. Wang, T., Yang, X., Ren, X., Zhao, J., Lam, K.Y.: Adaptive differentially private data stream publishing in spatio-temporal monitoring of IoT. In: 2019 IEEE 38th International Performance Computing and Communications Conference (IPCCC), pp. 1–8. IEEE (2019)
31. Wang, Y., Wu, X., Hu, D.: Using randomized response for differential privacy preserving data collection. In: EDBT/ICDT Workshops, vol. 1558 (2016)