Tao TAO, Funan ZHANG, Xiujun WANG, Xiao ZHENG, Xin ZHAO, 2024. An efficient online histogram publication method for data streams with local differential privacy. *Frontiers of Information Technology & Electronic Engineering*, 25(8):1096-1109. https://doi.org/10.1631/FITEE.2300368

An efficient online histogram publication method for data streams with local differential privacy

Key words: Data stream; Differential privacy; Sliding windows; Approximate counting

Corresponding author: Xiujun WANG
E-mail: wxj@mail.ustc.edu.cn
ORCID: https://orcid.org/0000-0002-8758-5763

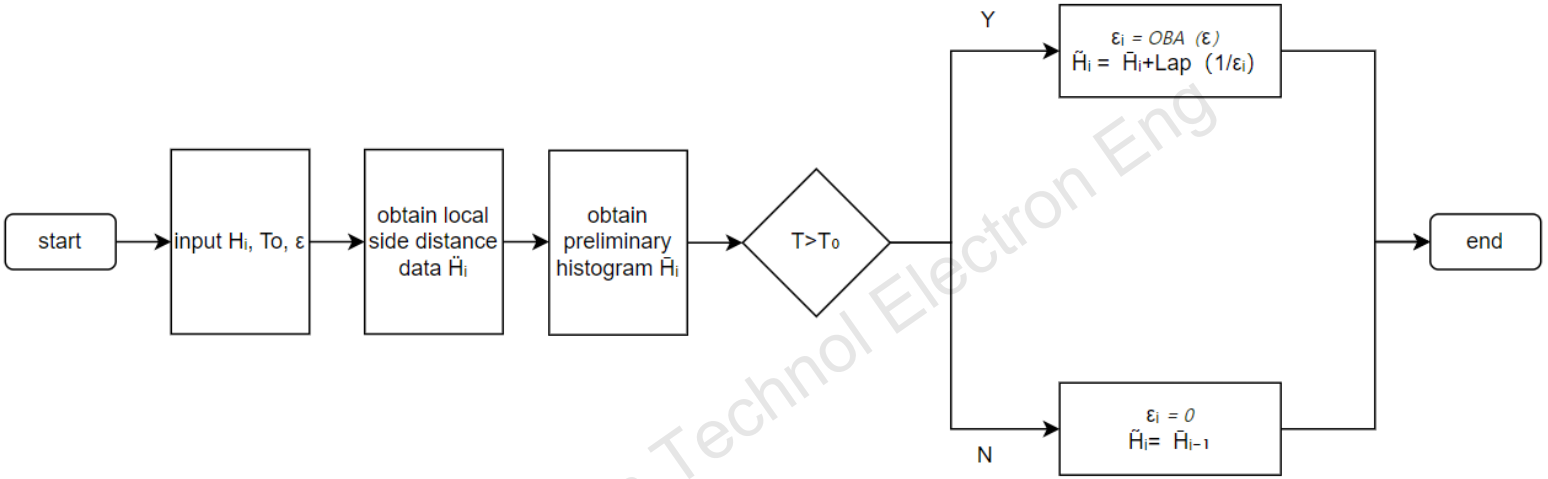
Motivation

- In the realm of data publishing, ensuring privacy while maintaining data utility is a fundamental challenge. Traditional approaches often fail to provide adequate privacy guarantees, particularly in the context of online data streams where data points arrive continuously over time. Local differential privacy (LDP) has emerged as a promising solution that allows individuals to share their data with untrusted servers without revealing sensitive information.
- However, existing LDP mechanisms do not fully address the complexities of data streams, leading to trade-off between privacy and utility. Our work aims to bridge this gap by developing novel LDP techniques specifically tailored for the publication of histograms from data streams.

Main idea

- To address the challenges of continuously generating publishable histograms from data streams while maintaining LDP, we design an efficient online histogram publication (EOHP) method. The EOHP method leverages an approximate counting technique to construct histograms of the current sliding window (SW) in an online manner, significantly reducing the time and storage costs compared to existing methods.
- We employ a scheme of adaptive noise enhancement and an optimized budget absorption (OBA) mechanism to carefully manage the privacy budget and add a suitable amount of noise to the histogram. This ensures that the histogram meets the privacy protection requirements while avoiding excessive consumption of the privacy budget. Such a design allows the EOHP algorithm to maintain a high level of data utility while providing robust privacy guarantees.
- Combining these two designs, we propose the EOHP algorithm and evaluate it on both synthetic and real-world benchmarks. The experimental results demonstrate that the EOHP algorithm outperforms state-of-the-art methods in terms of time and storage costs, as well as mean squared error (MSE), thereby achieving a better balance between privacy and data utility.

Framework



EOHP method for processing LDP data streams

It comprises three main steps: local data perturbation by users, rapid histogram generation for the current sliding window using an approximate counting method, and noise addition through an OBA mechanism to publish data that meets privacy protection requirements.

Method

Algorithm 2 Approximate counting algorithm

```
if len(B) > 0 and t − w == B[0][time] then
    del B[0]
end if
bit = f.readline()
if bit == 1/n then
    B = {time : i + 1, sum : 1}
    for i = len(B) − 1 to max − 1 do
        if B[i][sum] == B[i − max][sum] then
            B[i − max][sum] += B[i − max + 1][sum]
            B[i − max][time] = B[i − max][sum]
            del B[i − max + 1]
        end if
    end for
end if
if len(B) > 0 then
    for i = 0 to len(B) do
        sum + = B[i][sum]
        sum − = B[0][sum]/2
    end for
end if
return sum
```


To quickly obtain the statistical value of each interval in SW at the current time instance, the data collector uses the approximate counting algorithm to initially process the perturbed data, because this process only needs to cache the data in the current SW. This saves time and space costs significantly, and can improve the overall performance of the EOHP algorithm.

Method

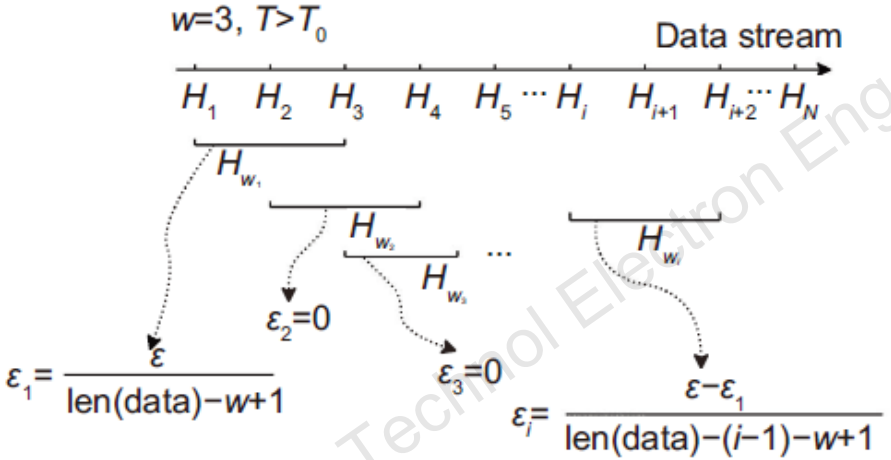


Fig. 2 Privacy budget allocation process of optimized budget absorption

To avoid over-exploitation of the privacy budget, we design an OBA mechanism that adds an appropriate amount of noise based on the error bound, to meet differential privacy (DP) protection requirements.

This approach allows maximization of privacy budget saving upfront, thus addressing the problem of excessive consumption of privacy budgets.

Results

Table 2 Comparison of histogram publication algorithms

Method	Design idea	Time cost	Storage cost	Mean squared error (MSE)
DDHP	Adding noise to the exact histogram and publishing	$O(w)$	$O(w\log_2 L)$	$1.5\times10^5-3.5\times10^5$
k-means-dp	Combining a MapReduce framework with k-means clustering to add noise to clustered data for publishing	$O(w)$	$O(w\log_2 L)$	$1.2\times10^5-2.6\times10^5$
CLSA-LDP	Adding noise to the exact histogram with random perturbation processing for release	$O(w)$	$O(w\log_2 L)$	$0.8\times10^5-1.7\times10^5$
EOHP	Adaptive noise enhancement and publishing of the approximate histogram online after random combustion	$O(L\log_2 w)$	$O(\frac{L}{r-1}\log_2 w)$	$0.6\times10^5-1.4\times10^5$

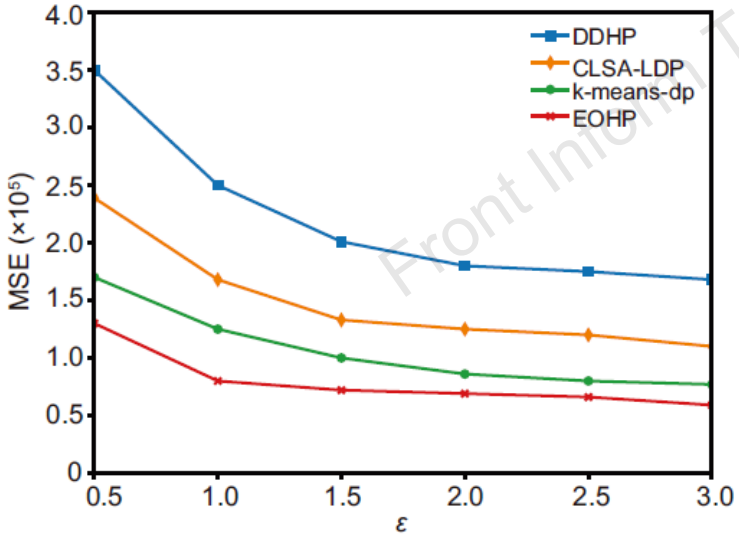


Fig. 4 Mean squared error (MSE) analysis of privacy budget ε on data utility based on the New York Taxi Dataset

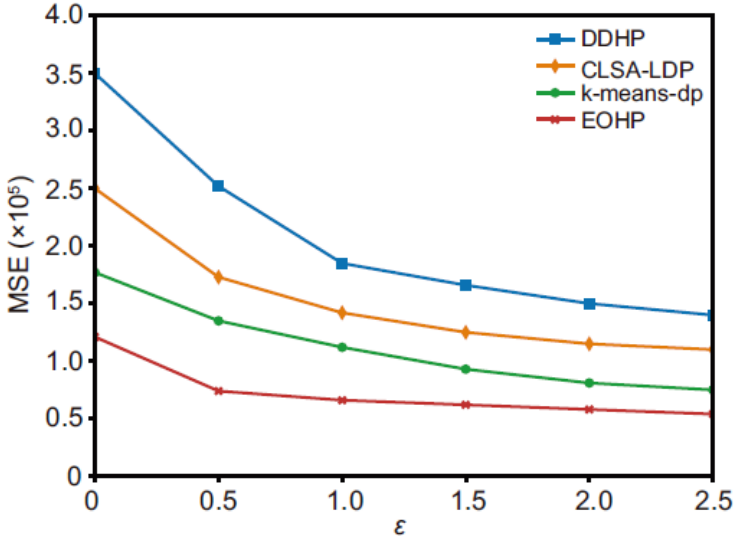


Fig. 5 Mean squared error (MSE) analysis of privacy budget ε on data utility based on the UK Car Accident Dataset

Results

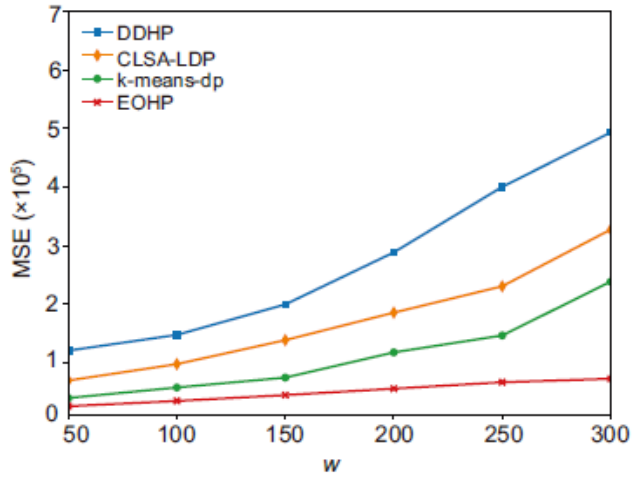


Fig. 6 Mean squared error (MSE) analysis of window size w on data utility based on the New York Taxi Dataset

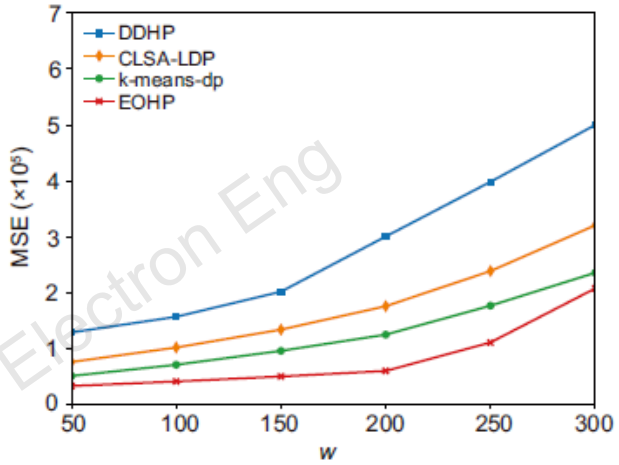


Fig. 7 Mean squared error (MSE) analysis of window size w on data utility based on the UK Car Accident Dataset

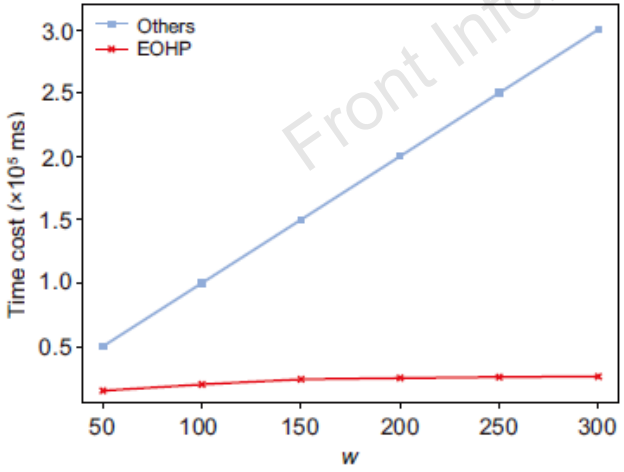


Fig. 8 Time cost comparison among EOHP, DDHP, k-means-dp, and CLSA-LDP algorithms

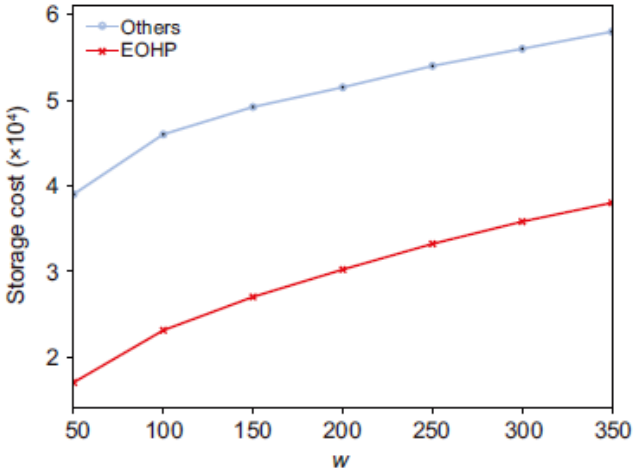


Fig. 9 Storage cost comparison among EOHP, DDHP, k-means-dp, and CLSA-LDP algorithms

Conclusions

In this paper, we introduce the EOHP algorithm for LDP data streams, enhancing privacy and efficiency. It employs random response for local data obfuscation, preventing privacy breaches. Utilizing an approximate counting method, it generates preliminary histograms online, reducing time and storage costs. The OBA mechanism ensures the optimal noise addition for privacy without exceeding the budget. Overall, EOHP outperforms the existing real-time data stream privacy algorithms, balancing protection with utility.

Tao TAO is currently a professor at the School of Computer Science and Technology, Anhui University of Technology. His research interests include intelligent Internet of Things, embedded systems, and data privacy protection.

Funan ZHANG received her MS degree in electronic information from Anhui University of Technology. Her research focuses on real-time data streams and data privacy protection.

Xiujun WANG received his BS, MS, and PhD degrees in computer science and technology from USTC. He is currently an associate professor at the School of Computer Science and Technology, Anhui University of Technology. His research interests include data stream processing and RFID system management.

Xiao ZHENG received his BS degree in computer software from Anhui University, his MS degree in engineering from the School of Information and Electronics at Zhejiang Sci-Tech University, and his PhD degree in computer application technology from Southeast University. His research interests include Industrial Internet, crowd-sensing network, cloud computing, and data privacy protection.

Xin ZHAO received her MS degree from Shandong Normal University. She is currently a teacher at Shengli No.1 Middle School of Dongying City, and her research interests mainly include cloud computing and database system management.