Locally Differential Private Succinct Sketches for Frequency Estimation

Hsuan-Po Liu, Rohan Putcha, and Haisu Li

***Abstract*—Frequency estimation in data streams is a foundational problem in computer science and data analysis. The goal is to efficiently determine the approximate frequency of items in data streams where storage and computational resources are limited. In this project, we propose a novel approach to frequency estimation that prioritizes privacy preservation by utilizing Count-Min Sketch (CMS) and local differential privacy (LDP) techniques for a distributed setting, i.e., an untrusted central server that aims to analyze the frequency of items according to multiple clients. Experiments are done on the proposed LDP-CMS to analyze the performance in terms of different privacy guarantee levels.**

I. Introduction

Frequency estimation in data streams is a foundational problem in computer science and data analysis, with applications spanning across real-world applications. The goal is to efficiently determine the approximate frequency of items in data streams where storage and computational resources are limited. Accurate frequency estimation is crucial for tasks such as detecting network anomalies, identifying trending topics on social media, or implementing privacy-preserving data analytics.

Count-Min Sketch (CMS) [1], [2] is a well-known probabilistic data structure for approximate frequency estimation in large-scale data streams. It is particularly useful when the dataset is too large to store or process in memory, providing a space-efficient way to estimate the frequency of elements in the data. In [1], CMS is defined as a two-dimensional data structure as a matrix $\mathbf{C}$ with size $d \times w$ and parameters (α, β), where $w = \lceil e/\alpha \rceil$ is the width and $d = \lceil \ln(1/\beta) \rceil$ is the depth, where α is the error factor and β is the failure probability. Each entry in $\mathbf{C}$ is initialized to 0. A set of d hash functions $\{h_1, \ldots, h_d\} : [n] \to [w]$ is chosen independently and randomly from a hash family, which maps a set of positive integers $\{1, ..., n\} \equiv [n]$ to $[w]$. While one might consider using raw data without approximation, in many practical scenarios, the memory required to store the full dataset exceeds the available resources. Thus, CMS offers a tradeoff: by allowing approximate frequency estimation with a significantly reduced memory footprint, it becomes an attractive solution for large-scale data streams. Following the professor's concern during our presentation, we formalize this in the appendix with a concise analysis of the CMS error bound and memory tradeoff, demonstrating under which conditions the CMS is an efficient alternative to storing raw data.

Differential Privacy (DP) is a privacy-preserving paradigm that enables data collection and analysis while safeguarding information. It ensures that the existence of any single individual's data in a dataset has a minimal impact on the output of an algorithm, thereby limiting the risk of revealing sensitive information. This is achieved by introducing carefully calibrated noise into the computation process, balancing privacy guarantees with the utility of the results. The increasing prevalence of data streams in various domains necessitates the development of efficient and privacy guaranteed tools, e.g., sketch-based analysis with DP guarantees [3]–[6]. Unlike central DP (CDP), which is impractical for assuming a trusted central server to introduce noise, local DP (LDP) operates in a distributed manner where users perturb their data locally before sharing it with the untrusted central server, which is our major focus in this project.

This project proposes a novel approach for frequency estimation, LDP-CMS, that prioritizes privacy preservation by utilizing CMS and LDP techniques for a distributed setting, i.e., an untrusted central server that aims to analyze the frequency of items according to multiple clients. Our approach draws inspiration from previous research that combines sketching techniques with differential privacy to balance accuracy, efficiency, and privacy. The proposed LDP-CMS are evaluated to analyze the mean squared error (MSE) of both CDP and LDP.

This report is for the term project in CSE 598: Why, What, and How.

II. LDP-CMS

In this section, we start with the problem setting of our project, followed by the privacy guarantees on streams. Then, we propose the method LDP-CMS.

A. Problem Setting

We consider a distributed system with N clients, each holding a local private stream. An untrusted central server collects information from all clients for frequency estimation. The stream held by client l is essentially a vector with t updates denoted by $\mathbf{s}_{[t]}^{(l)} = (s_1^{(l)}, \ldots, s_t^{(l)})$, where $s_1^{(l)}$ is the first event occurred and new events are appended as time elapses, $s_t^{(l)}$ is the event at timestamp t, $s_k^{(l)} \in \mathbb{Z}_n^+ = [n]$, for $k \in [t]$, and $l \in [N]$. In particular, clients $l \in [N]$ generate local count-min sketches $\mathbf{C}^{(l)}$'s with dimension $d \times w$ by local streams $\mathbf{s}_{[t]}^{(l)}$'s and publicly known randomly chosen d independent hash functions $\{h_1, \ldots, h_d\} : [n] \to [w]$, where we denote $\mathbf{C}^{(l)} = \mathrm{CMS}(\mathbf{s}_{[t]}^{(l)}; \{h_1, \ldots, h_d\})$. For ease of notations, we omit the set of hash functions to $\mathbf{C}^{(l)} = \mathrm{CMS}(\mathbf{s}_{[t]}^{(l)})$. The settings of the hash functions for the CMS follow [2] as $h_i(s_k^{(l)}) = ((a_i s_k^{(l)} + b_i) \mod p) \mod w$, where $a_i \neq 0$ and b_i are random integers, p is a prime number, for $i \in [d]$, $k \in [t]$, and $l \in [N]$. Then, clients $l \in [N]$ perturb $\mathbf{C}^{(l)}$'s by sampling noise matrices $\mathbf{Z}^{(l)}$'s with dimension $d \times w$, where

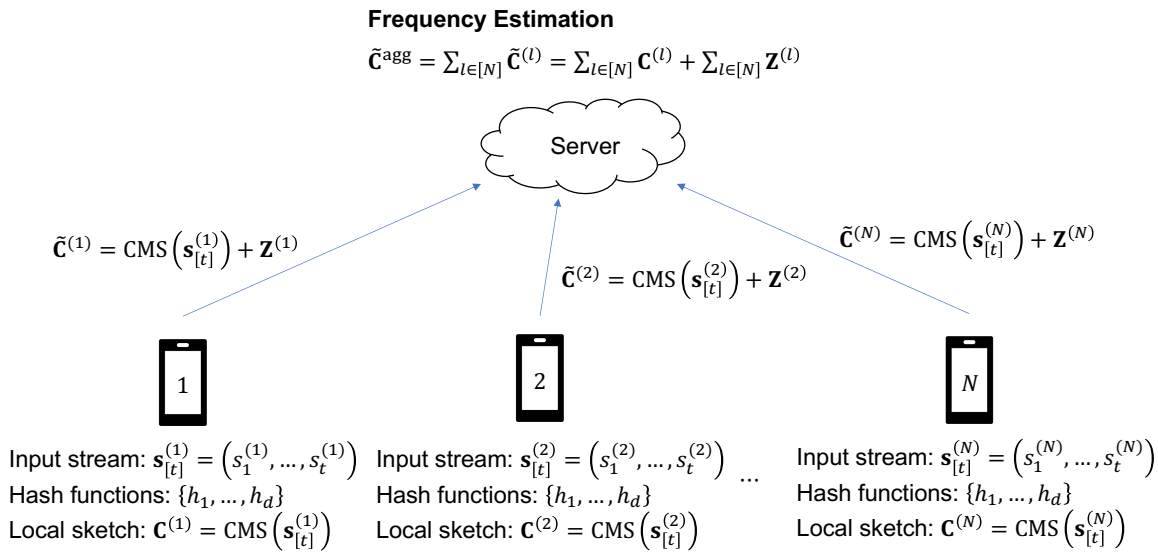


Fig. 1: Problem setting

each entry is sampled from $\mathcal{N}(0, \sigma^2)$, which is a zero mean Gaussian distribution with variance σ^2. We denote the perturbed local count-min sketch at client l as $\tilde{\mathbf{C}}^{(l)} = \mathbf{C}^{(l)} + \mathbf{Z}^{(l)}$, for $l \in [N]$. Lastly, clients $l \in [N]$ send $\tilde{\mathbf{C}}^{(l)}$'s to the server to aggregate $\tilde{\mathbf{C}}^{\text{agg}} = \sum_{l\in[N]} \tilde{\mathbf{C}}^{(l)} = \sum_{l\in[N]} \mathbf{C}^{(l)} + \mathbf{Z}^{(l)}$ for frequency estimation. The problem setting is summarized in Fig. 1.

B. Privacy Guarantee on Streams for CMS

Definition 1 ((ϵ, δ)-LDP on streams)**.** Let $\mathbf{s}_{[t]}^{(l)}$ and $\mathbf{s}_{[t]}^{(l)\prime}$ be neighboring streams, both have collected t events at timestamp t, in client l that differs only by a single event. The neighboring streams $\mathbf{s}_{[t]}^{(l)}$ and $\mathbf{s}_{[t]}^{(l)\prime}$ satisfy (ϵ, δ)-local differential privacy for any $\epsilon > 0$ and $\delta \in [0, 1]$ under a randomized mechanism $\mathcal{M}$ that under $\mathcal{E} \subseteq \text{Range}(\mathcal{M})$,

$$\mathbb{P}[\mathcal{M}(\mathbf{s}_{[t]}^{(l)}) \in \mathcal{E}] \leq e^{\epsilon} \cdot \mathbb{P}[\mathcal{M}(\mathbf{s}_{[t]}^{(l)\prime}) \in \mathcal{E}] + \delta, \tag{1}$$

where δ represents the failure probability, for $l \in [N]$.

Definition 2 (Sensitivity)**.** For two neighboring streams $\mathbf{s}_{[t]}^{(l)}$ and $\mathbf{s}_{[t]}^{(l)\prime}$ in client l together with CMS, the sensitivity is defined as follows:

$$\Delta \stackrel{\text{def}}{=} \max_{\mathbf{s}_{[t]}, \mathbf{s}'_{[t]}} \left\| \text{CMS}(\mathbf{s}_{[t]}^{(l)}) - \text{CMS}(\mathbf{s}_{[t]}^{(l)\prime}) \right\|_F, \tag{2}$$

where $\|\cdot\|_F$ represents the Frobenius norm, for $l \in [N]$.

Proposition 1. *Given two neighboring streams $\mathbf{s}_{[t]}^{(l)}$ and $\mathbf{s}_{[t]}^{(l)\prime}$ at client l, we have $\left\| \text{CMS}(\mathbf{s}_{[t]}^{(l)}) - \text{CMS}(\mathbf{s}_{[t]}^{(l)\prime}) \right\|_F \leq \sqrt{2d}$, for $l \in [N]$.*

Proof. For ease of notation, we denote $\mathbf{s}_{[t]}^{(l)}$ by $\mathbf{s}_{[t]}$ and $\mathbf{s}_{[t]}^{(l)\prime}$ by $\mathbf{s}'_{[t]}$, respectively. Thus, we have

$$\begin{aligned}
&\left\| \text{CMS}(\mathbf{s}_{[t]}) - \text{CMS}(\mathbf{s}'_{[t]}) \right\|_F \\
=&\left\| \text{CMS}\big((\mathbf{s}_{[t-1]}, s_t)\big) - \text{CMS}\big((\mathbf{s}_{[t-1]}, s'_t)\big) \right\|_F \\
\stackrel{(a)}{=}&\left\| \big(\text{CMS}(\mathbf{s}_{[t-1]}) + \text{CMS}(s_t)\big) - \big(\text{CMS}(\mathbf{s}_{[t-1]}) + \text{CMS}(s'_t)\big) \right\|_F, \\
=&\left\| \text{CMS}(s_t) - \text{CMS}(s'_t) \right\|_F \\
=&\sqrt{\sum_{i=1}^{d} \left\| \text{row}_i(\text{CMS}(s_t)) - \text{row}_i(\text{CMS}(s'_t)) \right\|_2^2} \leq \sqrt{2d}
\end{aligned} \tag{3}$$

where (a) is by the linearity of CMS, $\text{row}_i(\text{CMS}(\cdot))$ represents the ith row of $\text{CMS}(\cdot)$, and based on the construction of CMS, we obtain that

$$\begin{aligned}
&\left\| \text{row}_i(\text{CMS}(s_t)) - \text{row}_i(\text{CMS}(s'_t)) \right\|_2^2 \\
&= \begin{cases} 2 & \text{for} \quad h_i(s_t) \neq h_i(s'_t) \\ 0 & \text{for} \quad h_i(s_t) = h_i(s'_t) \end{cases},
\end{aligned} \tag{4}$$

for $i \in [d]$. The equality holds when $h_i(s_t) \neq h_i(s'_t)$ for all $i \in [d]$. □

Thus, according to Proposition 1, we have sensitivity $\Delta = \sqrt{2d}$. Please note that we prove it on our own.

C. Proposed Method

Each client applies an LDP-CMS as in Algorithm 1 to sketch the frequencies of elements in their local stream. The client then adds noise by Gaussian mechanism to the sketch locally to ensure (ϵ, δ)-LDP guarantee and sends the noisy sketch to an untrusted central server. The server aggregates these sketches to perform *frequency estimation* while ensuring individual privacy.

Theorem 2 (Classical Gaussian mechanism [7])**.** *The proposed LDP-CMS is (ϵ, δ)-LDP for $\sigma \geq \sqrt{2\ln(1.25/\delta)} \cdot \Delta/\epsilon$, where $\epsilon \in (0, 1)$.*

The classical Gaussian mechanism of Theorem 2 is efficient in determining the noise variance. However, it limits the privacy budget to $\epsilon \in (0, 1)$, and the bound is far from tight, leading to an unnecessary and more significant noise variance. In the LDP setting, protecting privacy already introduced a huge noise variance. Thus, we consider the analytic Gaussian mechanism [8] to tackle the problem.

Theorem 3 (Analytic Gaussian mechanism [8])**.** *For any $\epsilon > 0$ and $\delta \in [0, 1]$, the Gaussian mechanism is (ϵ, δ)-LDP if and only if $\Phi\big(\frac{\Delta}{2\sigma} - \frac{\epsilon\sigma}{\Delta}\big) - e^{\epsilon}\Phi\big(-\frac{\Delta}{2\sigma} - \frac{\epsilon\sigma}{\Delta}\big) \leq \delta$, where $\Phi(\cdot)$ represents the cumulative distribution function of the normal distribution.*

Theorem 3 optimizes the minimum value of σ^2 that guarantees (ϵ, δ)-LDP with $\epsilon > 0$. We denote it as a function that $\sigma^2 = \textsc{AnalyticGaussian}(\epsilon, \delta, \Delta)$. Thus, based on $\textsc{AnalyticGaussian}(\epsilon, \delta, \Delta)$, we propose the analytic Gaussian mechanism for LDP-CMS as follows.

Theorem 4 (Analytic Gaussian mechanism for LDP-CMS)**.** *Consider client l applying CMS to its stream $\mathbf{s}_{[t]}^{(l)}$, for $l \in [N]$. The Gaussian mechanism $\mathcal{M}$ to ensure (ϵ, δ)-LDP is defined as*

$$\mathcal{M}(\mathbf{s}_{[t]}) \stackrel{\text{def}}{=} \text{CMS}(\mathbf{s}_{[t]}) + \mathbf{Z}, \tag{5}$$

where $\mathbf{Z}$ is a noise matrix random noise to the query result according to $\mathcal{N}(0, \sigma^2)$, a zero-mean Gaussian distribution with variance $\sigma^2 = \textsc{AnalyticGaussian}(\epsilon, \delta, \Delta)$.

With Theorem 4, we summarize our LDP-CMS protocol in Algorithm 1.

Algorithm 1 LDP-CMS

Input: Privacy parameters ϵ, δ, Δ, streams from all clients $\mathbf{s}_{[t]}^{(l)} = [s_1^{(l)}, \ldots, s_t^{(l)}]$, for $l \in [N]$.

Server:

Choose d independent hash functions randomly $\{h_1, \ldots, h_d\} : [n] \to [w]$.

Share $\{h_1, \ldots, h_d\}$ to all clients.

Clients:

for clients $l \in [N]$ **do**

  Client l initialized the CMS $\mathbf{C}^{(l)}$ with a dimension of $d \times w$ such that all entries are 0's.

  for $k \in [t]$ **do**

    for $i \in [d]$ **do**

      $c^{(l)}_{i,h_i(s_k^{(l)})} += 1$ // Note: $c^{(l)}_{i,h_i(s_k^{(l)})}$ is the entry of index $(i, h_i(s_k^{(l)}))$ in $\mathbf{C}^{(l)}$

    end for

  end for

end for

Client l samples a noise matrix $\mathbf{Z}^{(l)}$ with dimension $d \times w$, where each entry is sampled from a zero mean Gaussian distribution with variance $\sigma^2 = \textsc{AnalyticGaussian}(\epsilon, \delta, \Delta)$.

Client l sends the noisy sketch $\tilde{\mathbf{C}}^{(l)} = \mathbf{C}^{(l)} + \mathbf{Z}^{(l)}$ to the server.

Server:

Aggregates $\tilde{\mathbf{C}}^{(l)}$'s to $\tilde{\mathbf{C}}^{\text{agg}} = \sum_{l\in[N]} \tilde{\mathbf{C}}^{(l)} = \sum_{l\in[N]} \mathbf{C}^{(l)} + \mathbf{Z}^{(l)}$.

Frequency estimation on $q \in [n]$ that $\hat{f}_q = \min_{i\in[d]} \tilde{c}^{\text{agg}}_{i,h_i(q)}$.

III. Implementation

In this section, we demonstrate our implementation based on the proposed LDP-CMS. We start by specifying the chosen parameters. Then, we visualize the results in terms of histograms and provide MSE results varying the privacy budget. Lastly, we discuss our observations based on the results.

A. Setup Parameters

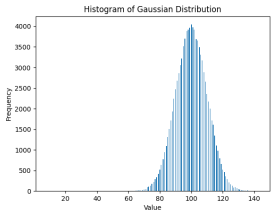
The following parameters were used to set up the LDP-CMS in our distributed system:

- **Number of Clients**: 5
- **Dataset Size**: 100,000 data points, with each client holding a stream of 20,000 data points.
- **Data Distribution**: Each event s_k is drawn from a normal distribution $\mathcal{N}(100, 100)$, rounded to the nearest integer and clipped to the domain $[1, 150]$.
- **Count-Min Sketch Parameters**:
  - Width $w = 50$
  - Depth $d = 10$
  - Hash Functions: Each hash function is of the form $h(x) = (ax + b \mod p) \mod w$, where:
    - p is a prime number sampled from the range $[2^{10}, 2^{18}]$.
    - a, b are uniformly sampled from $[1, p - 1]$.
- **Noise Parameters**:
  - $\delta = 10^{-3}$
  - Sensitivity $\Delta = \sqrt{2d} = \sqrt{20}$
  - σ^2 determined by the Analytic Gaussian mechanism over various values of ϵ.
- **Error Metric**: Mean Squared Error (MSE)

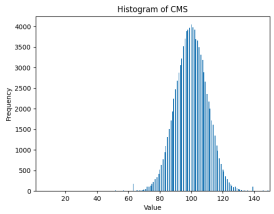
B. Histograms at Each Stage of Noise Application

We visualize the frequency distributions through histograms. Figure 2 demonstrates the histograms without DP guarantees. The true distribution of all items from all data streams is in Fig. 2a, while the CMS on the true items is in Fig. 2b with MSE $= 270.09$, small outliers can be observed due to error estimation.

Figure 3 presents the histograms with CDP and LDP guarantees under $\epsilon = \{0.5, 1, 10\}$. Figures 3a and 3b demonstrates the histogram with CDP and LDP under $\epsilon = 0.5$, yielding MSE $= 95500.63$ and MSE $= 404631.42$, respectively. Figures 3c and 3d demonstrates the histogram with CDP and LDP under $\epsilon = 1$, yielding MSE $= 9774.87$ and MSE $= 39311.42$, respectively. Figures 3e and 3f demonstrates the histogram with CDP and LDP under $\epsilon = 10$, yielding MSE $= 281.17$ and MSE $= 309.78$, respectively. In the figures, all CDP guarantees perform better than LDP-CMS, yielding better utility. However, LDP-CMS provides a stronger privacy guarantee and is generally more practical in real-world

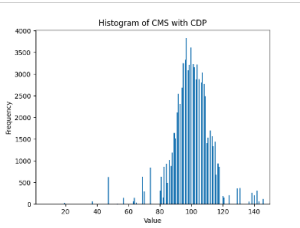


(a) True Histogram, No error

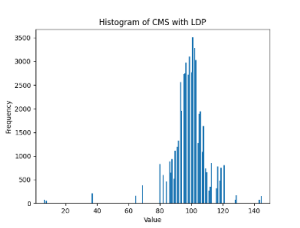


(b) CMS Error Histogram, MSE = 270.09

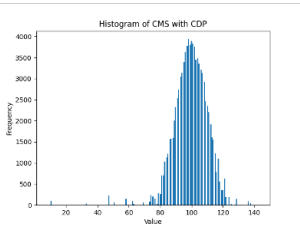
Fig. 2: Histograms of True Data and CMS Error Distribution



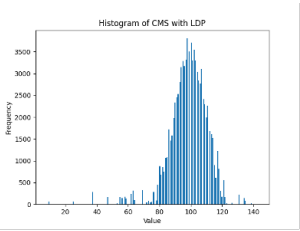
(a) CMS with CDP, $\epsilon = 0.5$, MSE = 95500.63



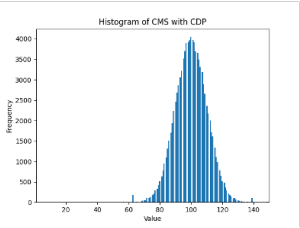
(b) CMS with LDP, $\epsilon = 0.5$, MSE = 404631.42



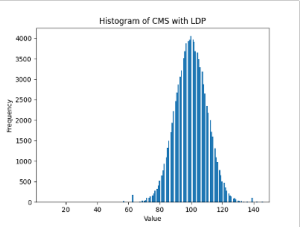
(c) CMS with CDP, $\epsilon = 1$, MSE = 9774.87



(d) CMS with LDP, $\epsilon = 1$, MSE = 39311.42



(e) CMS with CDP, $\epsilon = 10$, MSE = 281.17



(f) CMS with LDP, $\epsilon = 10$, MSE = 309.78

Fig. 3: Histograms After Applying CDP and LDP Noise with Varying ϵ Values.

applications. Also, the greater the ϵ, the better the utility with smaller MSE.

C. Mean Squared Error for Different ϵ Values

Figure 4 shows how the MSE changes for different privacy budgets ϵ under CDP and LDP settings. The error values decrease as ϵ increases, indicating improved accuracy. Table I summarizes the tested ϵ values along with the corresponding parameters for δ, Δ, and σ^2.

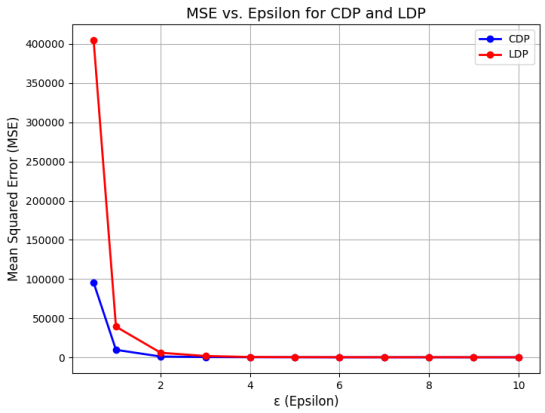


Fig. 4: MSE vs. ϵ for CDP and LDP

D. Key Observations

Based on the experimental results, we summarize the following observations:

- **Error Reduction**: As ϵ increases, the MSE decreases, indicating better accuracy. The MSE approaches the baseline CMS error (MSE = 270.09 in our implementation).
- **CDP vs. LDP**: CDP consistently outperforms LDP in accuracy, regardless of the ϵ value.
- **Utility vs. Privacy Trade-off**: Higher ϵ provides better utility but reduces privacy guarantees, demonstrating the typical trade-off between accuracy and privacy.

For further details and code, please refer to the GitHub Repository.

IV. Conclusion

In this project, we have proposed that LDP-CMS tackle the privacy guarantees for CMS in a distributed setting. The privacy guarantee for LDP on stream is defined. We use the analytic Gaussian mechanism to optimize the noise variance to achieve better utility while preserving the same privacy level. The LDP-CMS is implemented to evaluate MSE performance. We leave exploring cryptographic methods as a complementary enhancement to our core framework, potentially enabling more robust privacy guarantees and higher utility as future work.

Appendix

A. Count-Min Sketch (CMS) error bound and memory tradeoff

The Count-Min Sketch (CMS) is a probabilistic data structure used to approximate the frequency of items in a data stream, significantly reducing memory usage compared to storing the full dataset. The CMS consists of a matrix with d rows and w columns, where d is the depth (number of hash functions) and w is the width (number of columns per row).

The CMS error bound is:

$$\Pr(\tilde{f} \leq f + \alpha t) \geq 1 - \beta$$

where $\tilde{f}$ is the estimated frequency, f is the true frequency, α is the error factor, t is the number of items in the stream, and β is the failure probability.

ϵ	0.5	1	2	3	4	5	6	7	8	9	10
δ	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}	10^{-3}
Δ	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$	$\sqrt{20}$
σ^2	425.07	132.57	41.77	21.52	13.55	9.52	7.16	5.65	4.61	3.86	3.29

TABLE I: Parameters for Different ϵ Values

For a given error bound α and failure probability β, the standard formulas for w and d are:

$$w = \left\lceil \frac{e}{\alpha} \right\rceil, \quad d = \left\lceil \ln(\frac{1}{\beta}) \right\rceil$$

The asymptotic memory usage of the CMS is as follows:

$$\text{Memory} = O(w \times d)$$

To be efficient, $w \times d$ should be much smaller than the total number of stream items t, i.e., $w \times d \ll t$.

There is a trade-off between memory and accuracy: reducing w and d lowers memory usage but increases error α, while increasing w and d improves accuracy but requires more memory. Thus, the CMS is only ideal to use with some approximation error α if $w \times d$ takes up far less memory than sending data without the CMS.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to Professor Roy Chowdhury for her kind support and guidance throughout this project. Her willingness to meet with us in person during office hours was instrumental in helping us navigate challenges and refine our work.

REFERENCES

[1] G. Cormode and S. Muthukrishnan, "An improved data stream summary: the count-min sketch and its applications," *Journal of Algorithms*, vol. 55, no. 1, pp. 58–75, 2005.

[2] L. Melis, G. Danezis, and E. De Cristofaro, "Efficient private statistics with succinct sketches," *arXiv preprint arXiv:1508.06110*, 2015.

[3] M. Aumüller, A. Bourgeat, and J. Schmurr, "Differentially private sketches for jaccard similarity estimation," in *Similarity Search and Applications: 13th International Conference, SISAP 2020*, 2020, pp. 18–32.

[4] A. Smith, S. Song, and A. Guha Thakurta, "The flajolet-martin sketch itself preserves differential privacy: Private counting with minimal space," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 561–19 572, 2020.

[5] R. Pagh and M. Thorup, "Improved utility analysis of private countsketch," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 631–25 643, 2022.

[6] F. Zhao, D. Qiao, R. Redberg, D. Agrawal, A. El Abbadi, and Y.-X. Wang, "Differentially private linear sketches: Efficient implementations and applications," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 691–12 704, 2022.

[7] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[8] B. Balle and Y.-X. Wang, "Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *International Conference on Machine Learning*. PMLR, 2018, pp. 394–403.