A Universal Sketch for Estimating Heavy Hitters and Per-Element Frequency Moments in Data Streams with Bounded Deletions

Anonymous Author(s)

ABSTRACT

In the field of data stream processing, there are two prevalent models, i.e., insertion-only, and turnstile models. Most previous works were proposed for the insertion-only model, which assumes new elements arrive continuously as a stream, and neglects the possibilities of removing existing elements. In this paper, we make a *bounded deletion* assumption, putting a constraint on the number of deletions allowed. For such a turnstile stream, we focus on a new problem of *universal measurement* that estimates multiple kinds of statistical metrics simultaneously using limited memory and in an online fashion, including per-element frequency, heavy hitters, frequency moments, and frequency distribution. There are two key challenges for processing a turnstile stream with bounded deletions. Firstly, most previous methods for detecting heavy hitters cannot ensure a bounded detection error when there are deletion events. Secondly, there is still no prior work to estimate the per-element frequency moments under turnstile model, especially in an online fashion. In this paper, we address the former challenge by proposing a Removable Augmented Sketch, and address the latter by a Removable Universal Sketch, enhanced with an Online Moment Estimator. In addition, we improve the accuracy of frequency estimation by a compressed counter design, which can halve the memory cost of a frequency counter and support addition/minus operations. Our experiments show that our solution outperforms other algorithms by 16% ∼ 69% in F1 Score of heavy hitter detection, and improves the throughput of frequency moment estimation by 3.0×10^4 times.

CCS CONCEPTS

• **Theory of computation → Sketching and sampling**.

KEYWORDS

Data streams, Turnstile Model, Sketch, Universal Measurement

ACM Reference Format:

Anonymous Author(s). 2018. A Universal Sketch for Estimating Heavy Hitters and Per-Element Frequency Moments in Data Streams with Bounded Deletions. In *Proceedings of ACM Manag. Data (SIGMOD).* ACM, New York, NY, USA, 24 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Background. In many big data scenarios, data arrives as a continuous stream and at a high speed, such as Internet traffic logs [2, 11, 17, 27, 42, 59, 76, 77, 86], sensor network readings [14, 35, 61, 83, 91], and social media messages [4, 12, 40, 49, 52, 54]. Extensive prior studies have been devoted to designing time-efficient one-pass algorithms for data streams. For example, in network traffic measurement, AT&T collects terabytes of NetFlow [25] data from its production network every day. Numerous network flow logs constitute a data stream, which is processed by one pass to extract valuable statistical information for monitoring IP network services [24].

Universal Sketch. Many kinds of statistical metrics can be measured for a data stream. The first measurement task is to estimate the per-element frequency [16, 72, 76, 94], e.g., counting the number of packets originated from each source IP. The second task is to identify ε-heavy hitters, i.e., the elements whose frequencies exceed ε percent of the total frequency [10, 16, 42]. The third task is to measure the aggregated information of all elements, called the moments of element frequencies [20, 26, 73, 78]. For instance, the 0th-order moment is the number of distinct elements. The fourth task is to reconstruct the distribution of per-element frequencies, which can be achieved by fitting the generic moments [5, 23, 68, 89].

However, most existing algorithms are proposed to estimate only one type of the above statistics [10, 16, 19, 33, 55, 74, 93]. If each type of metric is measured by one dedicated algorithm, it will consume extensive memory and computation resources. Hence, *universal sketches* have been proposed to measure multiple metrics by one algorithm [42, 72, 76, 83]. They are developed based on a *recursive summation technique* to estimate generic moments [7]. This technique uses a two-phase framework: (a) In the online updating phase, the data stream is recursively sampled into multiple substreams by different probabilities $1, \frac{1}{2}, \frac{1}{4}, \ldots$. For each substream, a *subsketch* is utilized to capture the IDs of heavy hitters, and their frequencies. (b) In the offline estimation phase, the moment of each sampled substream is calculated recursively from the captured heavy hitters.

Bounded Deletion Stream. The previous universal sketches [42, 72, 83, 89] are designed to process an *insertion-only* stream. As a result, when deletion events arrive, they are unable to correspondingly update the multiple statistics. In real-world scenarios, there are two prevalent data stream models, namely the *insertion-only model* and the *turnstile model*. An *insertion-only* stream contains only insertion events, i.e., element IDs with frequency increments. A *turnstile* stream contains both insertion events and deletion events, i.e., frequency decrements of existing element IDs. For example, a proportion of NetFlow records may be deleted by network administrators, since they are collected during network attack campaigns.

In this paper, we focus on a new problem of designing a universal sketch that scans a *turnstile* stream by one pass and estimates the above statistics in an online manner. For the *turnstile* stream, a plausible assumption is that all the past insertions can be deleted. However, this incurs dramatically higher memory cost and longer update time than those of the *insertion-only* stream [8, 70, 90].

SIGMOD, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

According to Jayaram et al. [29], for many *turnstile* data streams in practice, there are only a small fraction of element deletions. For example, the NetFlow logs deleted by network administrators due to their relation with network attacks only occupy a small fraction in a network log database. As a result, they proposed the *bounded deletion model* to ensure the number of deletions does not exceed $1 - \frac{1}{\zeta}$ fraction of the number of prior insertions, where ζ is a pre-configured constant no smaller than 1. When ζ is infinity, all past inserted elements can be deleted, and detecting heavy hitters needs significant memory cost linear to the massive number of elements. As ζ decreases and approaches 1, the memory cost can be greatly reduced [6, 29, 90, 92]. In this paper, we measure multiple kinds of statistics under the *bounded deletion model* and in an online manner.

Motivation. For generic measurement of a data stream, researchers have proposed the framework of universal sketch [42, 73]. As shown in Fig. 1, it consists of multiple components: substream sampler, heavy hitter (HH) detector for each substream, generic moment estimator, and the distribution fitter. Among them, the most important components are the HH detector and the moment estimator. These components face the following challenges when online processing a *bounded deletion* stream with millions or even billions of elements.

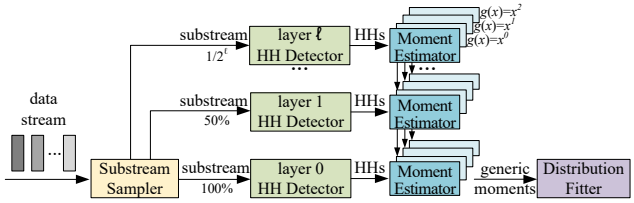


Figure 1: Components of a general universal sketch

Challenge 1 (HH Detection in the Bounded Deletion Model). As shown in Fig. 1, a HH detector is to process a substream and identify the elements whose frequencies exceed a threshold. However, to detect the HHs in a bounded deletion substream, existing methods adopt a *simple structure* with a hash table to hold HHs' element-frequency pairs [48]. Therefore, when an insertion/deletion event arrives carrying an element ID untracked by the hash table, an existing element has to be chosen as a victim for replacement. This inevitably causes information loss, since the victim element has to be evicted from the table. So we design a *composite structure* combining the hash table with a backend data sketch to hold the non-HHs.

Challenge 2 (Improvement of HH Detection Accuracy). Traditionally, the backend data sketch uses a matrix of frequency counters, and each counter is implemented by a 32-bit integer. To improve the counters' memory efficiency, several recent studies propose a compression technique that implements each counter with 16-bit memory [83, 94]. This can double the number of counters in the sketch under the same memory budget, thereby improving HH detection accuracy. However, the previous counter designs are still inadequate. They either have a small counting range [83], e.g., $(-2^{15}, +2^{15})$, or lack support for the subtraction operation [72, 76], which is necessary for a turnstile stream with element deletions.

Challenge 3 (Online Moment Estimation). As shown in Fig. 1, for each substream, a HH detector is used to sample the heaviest elements. A moment estimator is also deployed to take a set of HHs as input, and estimate the moments of the frequencies of all elements in the substream. Previous work uses a *recursive summation technique* for moment estimation [7], which however has high computational cost and must be performed in an offline manner. This technique, as shown in Fig. 1, calculates the moment estimation by scanning the HHs of all substreams, starting from the ℓth sampling layer downwards to the 0th layer with 100% sampling probability.

Our Approach. In this paper, we propose a Removable Universal Sketch (RUS), which incorporates multiple improved designs.

To address **Challenge 1**, we design a *composite data structure* named Removable Augmented Sketch (RAS) that combines a pre-filtering *element-frequency table* and a backend sketch. The table holds both the IDs and the frequencies of heavy hitters. The backend sketch does not store the element IDs, but holds the frequencies of all elements in a compressed manner. The main advantage of this composite structure is that, when insertion or deletion events arrive carrying the elements untracked by the table, the backend sketch can update the frequencies of these untracked elements.

To address **Challenge 2**, we propose a novel compressed counter design, called the Removable Active Counter (RAC). RAC has a low memory cost with only 16 bits. It has a sign bit, an exponent part and a coefficient part, providing a wide counting range of $(-2^{27}, +2^{27})$. It outperforms other designs that halve the memory footprint at the cost of significantly narrowing the counting range [83]. Furthermore, our RAC supports both addition and subtraction operations.

To address **Challenge 3**, we propose the Online Moment Estimator (OME). It adapts the *recursive summation technique* [7], so that the generic moment estimation can be updated incrementally when the set of heavy hitters changes. Since the generic moments can be estimated online, it is possible to reconstruct the element frequency distribution online using the method of moments [5, 89].

In summary, we propose RUS to collect multiple statistics online in the bounded deletion model (Section 5). Our contributions are:

- We propose a composite data structure named RAS to track HHs in the bounded deletion model with high accuracy (Section 6).
- We propose a 16-bit RAC, which supports both addition and subtraction operations, and has a wide counting range (Section 7).
- We propose the OME, which allows us to update the generic moment estimation incrementally in an online manner (Section 8).
- We conduct extensive experiments (Section 9). The results show that our designs improve the F1 Score of HH detection by 16% ∼ 69%, and increase the moment estimation speed by about 3×10^4 times.

2 BACKGROUND

This section presents the data stream models and the problems we study. We summarize the notations commonly used in Table 1.

2.1 Data Stream Models

Formally, a data stream $S = \langle (e_1, w_1), (e_2, w_2), \ldots, (e_t, w_t) \rangle$ with a current length of t is a sequence of tuples, where each e_i indicates an element ID, and w_i represents the weight of e_i. Let E be the universe set of element IDs and we have $e_i \in E$. An element ID is allowed to appear multiple times in a data stream, which means the condition as $e_i = e_j$ with $i \neq j$ may occur. Depending on the domain of w_i, data streams can be classified into two major categories [50].

- **Insertion-only model** assumes all w_i are positive, which implies that the frequency of an element e_i can only increase and not decrease. This model is the most commonly used model.

• **Turnstile model** allows w_i to be positive or negative, for a tuple (e_i, w_i). It is an insertion if $w_i > 0$, and a deletion if $w_i < 0$. So the frequency of an element can both increase and decrease.

We assume a **bounded deletion model** with an upper-bounded $D:I$ ratio, which is the number of deletions D divided by the number of insertions I. More formally, $D:I \leq 1 - \frac{1}{\zeta}$, where ζ is a pre-configured constant no smaller than 1. This model definition is flexible. When $\zeta = \infty$, this model degrades to the *turnstile model* that allows all previously inserted elements to be deleted. When ζ is close to 1, the $D:I$ ratio is bounded, and the memory and time costs for detecting the ε-heavy hitters can be greatly reduced [6, 29, 90, 92].

2.2 Problem Definition

In this paper, we aim to measure multiple kinds of data stream statistics within a bounded deletion stream setting and in an online manner. We define these statistics as follows.

Per-Element Frequency. Let f_e be the frequency of an element e. When a tuple (e, w) arrives, we have $f_e = f_e + w$. Thus $f_e = \sum_{e_i=e} w_i$. Let $F = \sum_{e \in E} f_e$ be the total frequency of all elements in the set E.

Frequency Moment. The g-moment of the stream is the functional sum of the frequency f_e for all the elements e with arrival tuples:

$$L = \sum_{e \in E} g(f_e), \tag{1}$$

where $g(x)$ is a monotonic function bounded by x^2. Typical definitions of the function g are given as follows.

• If $g(x) = x^0 = 1$, then L is called the 0th-order moment. It is equal to the cardinality, i.e., the number of distinct elements $|E|$.
• If $g(x) = x$, then L is called the 1st-order moment. It is equal to the total frequency of all elements, namely, L equals $\sum_{e \in E} f_e$.
• If $g(x) = x \log x$, then L is the entropy of the frequencies of all elements, indicating the diversity of the frequency distribution.
• If $g(x) = x^2$, then L is the 2nd-order moment of the frequencies of all elements, indicating the variance of the frequency distribution.

Heavy Hitters. A heavy hitter is an element $e \in E$, whose frequency f_e contributes at least ε fraction of the g-moment $L = \sum_{e \in E} g(f_e)$. We denote the set of all heavy hitters H as Eq. (2), where $0 < \varepsilon < 1$.

$$H = \{e \mid g(f_e) \geq \varepsilon L\} \tag{2}$$

We call H the first-order heavy hitters or $L1$ heavy hitters if $g(x) = x$, and second-order heavy hitters or $L2$ heavy hitters if $g(x) = x^2$.

Per-Element Frequency Distribution. The distribution considers the frequency of an element as a random variable. In the following paper, we refer to it as frequency distribution for brevity. The probability mass function of the distribution can be written as:

$$d(f^*) = \begin{cases} Pr(f_e = f^*), & f^* \in \mathcal{S}_f, \\ 0, & f^* \notin \mathcal{S}_f, \end{cases} \tag{3}$$

where $\mathcal{S}_f$ is the set of all possible per-element frequency values of a stream, and f^* is a random frequency value. $Pr(f_e = f^*)$ is the probability of an arbitrary element e to have a frequency value f^*.

3 RELATED WORK

In this section, we review the related work on per-element frequency estimation, heavy hitter detection, and the universal sketch.

Table 1: Symbols frequently used in this paper

Symbol	Description
S, S_j	Stream and substream sampled with probability $1 / 2^j$
(e, w)	A stream tuple with element ID e and weight w
f_e^{new}, f_e^{old}	The new frequency and old frequency of the element e
I, D	Number of insertion events and number of deletion events
ζ	The parameter to bound # deletions / # insertions (i.e., $D:I$)
L, L_j	Frequency moment of stream S and substream S_j
ε	The threshold to determine a heavy hitter
H, H_j	Set of all heavy hitters in stream S and substream S_j
C, ρ, α, β	Counter and its sign, exponent, and coefficient part
$Maxkicks$	The maximum number of kicks for KP-CF.

3.1 Per-Element Frequency Estimation

A data sketch is a technique to create compact summaries of large data streams [15]. It can be used to estimate per-element frequencies in a data stream, such as CountSketch (CS) [10], randomized error-reduction Sketch (rSkt) [66, 67], CountMin Sketch (CMS) [16], and Conservative Update Sketch (CUS) [19]. It uses hash functions to project the per-element frequency vector into a fixed-size structure with a sublinear memory cost to the number of distinct elements, allowing for efficient querying of an arbitrary element's frequency.

Most sketches allocate a matrix of 32-bit integers with d rows, where d is often set to 4. CMS (or CUS) can only provide an over-estimation, since for each tuple (e, w), it updates by adding the weight w to all (or the minimum) of the hashed counters. By contrast, CS provides an unbiased estimation using the Tug-of-War technique [3]. Specifically, before adding w to the hashed counter in each of the d rows, CS multiplies w randomly by either +1 or −1. rSkt also offers unbiased estimation. To avoid multiplication of +1 or −1, it uses two rows to implement the function of one row in CS.

Counter compression techniques boost sketch accuracy by allowing more counters to be allocated. Some of them involve altering the counting strategy. Diamond Sketch (DS) [84] uses 4-bit counters and utilizes a counting strategy similar to hexadecimal. FCM-Sketch (FCMS) [60] and TowerSketch [82] employ a counting strategy similar to DS but allocate different-sized counters in different rows. These strategies only work in the *insertion-only model*, thus DS allocates an extra CUS to record deletions. Others design counters for seamless integration into various sketches. The lossless encoding counter [83] uses integers in different sizes to represent numbers within various ranges, all of which are within -2^{15} to $+2^{15}$. Active Counter [47, 94] has a large counting range. ActiveCM [76] and GenericCM [76] use 16-bit active counters with a counting range of 0 to 2^{43} to enhance CMS. However, they do not support subtraction.

3.2 Heavy Hitter Detection

Existing approaches for identifying heavy hitters (HHs) can be categorized into counter-based and sketch-based algorithms [39].

Counter-Based Algorithms. These algorithms maintain an *element-frequency table* to hold the IDs and frequencies of the current HHs in a data stream. Canonical algorithms include SpaceSaving (SS) [48], Unbiased SpaceSaving (USS) [64], SpaceSaving$^{\pm}$ (SS$^{\pm}$) [90], Lossy Counting [46], HeavyGuardian [85], HeavyKeeper [87], WavingSketch [37], and Cuckoo Filter along Kicking Path (KP-CF) [75, 77]. The KP-CF provides state-of-the-art update throughput and estimation accuracy. However, it only works in the *insertion-only* stream.

Likewise, most existing methods, such as SS and USS, do not apply to the *bounded deletion model* due to Challenge 1 in Section 1.

SS$^{\pm}$ is an initial solution to handle bounded deletion streams [90]. When each tuple (e, w) arrives, e may not be cached by the table. If it is an insertion with $w > 0$, SS$^{\pm}$ adopts the same replacement strategy as SS. If it is a deletion with $w < 0$, SS$^{\pm}$ applies w to the element with the largest estimation error in the table. SS$^{\pm}$ proves that, if the capacity k of the table exceeds $\frac{\zeta}{\epsilon}$, the ϵ-heavy hitter detection error can be bounded by a threshold with a high probability.

We give an example in Fig. 2, where an *element-frequency-error table* stores the elements e_1, e_2, and e_3. When an insertion event $(e_4, 2)$ arrives carrying an untracked element e_4, SS$^{\pm}$ finds the record $(e_1, 7, 0)$ with the smallest frequency. Then, the record becomes $(e_4, 9, 7)$, as its element ID is replaced by e_4, its frequency is increased by 2, and the estimation error grows to 7 due to the replacement. Next, when a deletion event $(e_1, -1)$ arrives carrying an untracked element e_1, SS$^{\pm}$ finds the record $(e_4, 9, 7)$ with the largest error, and reduces its frequency and error by 1, resulting in $(e_4, 8, 6)$.

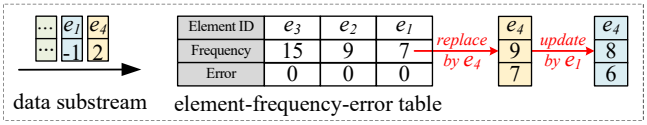


Figure 2: Example of handling insertion and deletion events

However, the previous works, including SS$^{\pm}$ sketch [90], have suboptimal estimation accuracy of per-element frequency, which in turn greatly reduces the identification accuracy of HHs. The main cause is that SS$^{\pm}$ adopts a simple data structure design with *one single element-frequency table* as shown in Fig. 2. Therefore, when an insertion or deletion event arrives carrying an element ID untracked by the table, it has to choose an existing element for replacement, which inevitably causes accuracy loss. For example, in Fig. 2, the frequency of element e_4 is 2, but is overestimated as 8.

Sketch-Based Algorithms. These algorithms maintain a frequency counting sketch, as well as a HH sampling structure guided by the sketch. Traditionally, a min-heap is placed after the sketch to hold the current top-k HHs [76]. We call it a postfilter. However, this solution does not fully utilize the highly skewed distribution of element frequencies, thus resulting in suboptimal accuracy. Recent works propose to separate HHs from normal elements by a prefilter-based strategy [55, 72, 79]. These algorithms maintain a pre-filtering *element-frequency table* to track current HHs, and a backend sketch to hold the evicted non-HHs, which greatly improves the HH detection accuracy by gracefully swapping out the non-HHs.

This strategy was first proposed by Augmented Sketch (AS) [55]. AS implements the prefilter by a min-heap, with $O(k)$ lookup time cost when holding the top-k HHs. However, for a turnstile stream, the size of the prefilter might need to be set to tens of thousands to ensure accuracy. Several follow-up works propose to improve the throughput by hash-based prefilter structures with $O(1)$ time cost, such as Cold filter [93], ElasticSketch [86], MV-Sketch [63], and OneSketch [22]. However, they only work in *insertion-only streams*.

3.3 Universal Sketch

Universal sketch can estimate multiple kinds of statistical metrics simultaneously, such as per-element frequency, heavy hitters, and generic moments. UnivMon [42], Light-weight Universal Sketch (LUS) [76], Augmented LUS (ALUS) [72], and Joltik [83] can natively support generic moment estimation, which relies on the *recursive summation technique* [7]. We neglect ElasticSketch [86], FCMS [60], OneSketch [22], and Panakos [91], which can only estimate lower-order moments (e.g., cardinality and entropy) but not higher-order moments [32, 69]. Higher-order moments are essential for precisely recovering the per-element frequency distribution [23, 30, 56].

UnivMon is the first universal sketch. It works in the *insertion-only model* and uses the *hierarchical sampling* method to favor the heavy hitters in the long tail. Follow-up studies propose better moment estimators based on different optimization techniques. Both LUS [76] and Joltik [83] replace the *hierarchical sampling* method with the *progressive sampling technique*. This technique dramatically improves the time efficiency and estimation accuracy. To reduce the memory footprint, the previously mentioned ActiveCM [76] and lossless encoding counter [83] are employed in LUS and Joltik, respectively. ALUS utilizes a prefilter [55] to reduce the estimation error of LUS. Overall, the universal sketch algorithms have achieved significant progress. However, none of these algorithms can estimate moments online in the *bounded deletion model*.

In summary, we compare our work with existing methods in Table 2. It shows that our work has two key advantages: the ability to handle deletion events, and online estimation of generic moments.

Table 2: Comparison of existing methods with our solutions

Task	Method	Accuracy	Throughput	Deletion Support	Moment Estimation
Per-Element Frequency Estimation	Traditional Sketches	○	●	●	–
	w. Compressed Counter	●	○	○	–
	Sketches with Our RAC	●	◐	●	–
Heavy Hitter Detection	Counter-Based	○	●	◐	–
	Sketch-Based	●	○	○	–
	Our RAS	●	●	●	–
Moment Estimation	UnivMon Inspired	◐	◐	○	Offline
	Our RUS with OME	●	●	●	Online

●, ◐, and ○ represent perform well, moderately, and poorly, respectively.

4 PRELIMINARY

In this section, we first give an introduction to the universal sketch. Then, we present the design of KP-CF to track heavy hitters (HHs).

4.1 Preliminary about Universal Sketch

The universal sketch is constructed as a layered structure, and employs the two-phase framework from *recursive summation* [7] to estimate moments. We take UnivMon as an example to illustrate the layered structure and the two-phase framework. Similar to Fig. 1, we assume UnivMon has multiple layers, whose indices range from 0 to ℓ. In each of the $\ell + 1$ layers, a CS combined with a postfilter is allocated to track heavy hitters, i.e., the *HH Detector* in Fig. 1, and we denote the combination in each jth layer as CS_j for simplicity.

Online Updating Phase. For an arrival tuple (e, w), the *hierarchical sampling* is performed, with a sampling probability $\frac{1}{2^j}$ assigned for the jth layer (as depicted in Fig. 1). Specifically, the element e is initially sampled at the 0th layer with a probability of 100%, then at the 1st layer with a probability of 50%, and the procedure continues until sampling fails at a particular layer. To clarify, the

element is sampled for all subsequent layers from the 0th layer up to the topmost sampled layer $j_t(e)$, as determined by Eq. (4).

$$j_t(e) = \min\left(\ell,\ \arg\max_j \left\{ \bigwedge\nolimits_{0\le i\le j} [1|h(e)]_i = 1 \right\}\right), \tag{4}$$

where $h(e)$ represents the binary representation of e's hash, and $[1|h(e)]_i$ is the leftmost ith bit of the concatenation of 1 with $h(e)$. Hence, the expression $\wedge_{0\le i\le j}[1|h(e)]_i = 1$ implies that the leftmost j bits of $[1|h(e)]$ are all 1s. Since each bit of $h(e)$ has 50% chance of being 1, the sampling probability reduces by half with each layer, and it is 100% for the 0th layer as $[1|h(e)]_0$ is always 1.

After that, among the sampled substream S_j on the jth layer, $j \in [0, j_t(e)]$, the sketch CS_j identifies the heavy hitters H_j in Eq. (2).

Offline Estimation Phase. Firstly, the g-moment L_ℓ of the substream S_ℓ sampled for the highest layer ℓ is estimated by

$$\hat{L_\ell} = \sum\nolimits_{e\in\hat{H_\ell}} g(\hat{f_e}). \tag{5}$$

Then, from layer $\ell-1$ to 0, the *recursive summation technique* [7] defined in Eq. (6) is used to obtain the g-moment of the substream S_0:

$$\hat{L_j} = 2\,\hat{L}_{j+1} + \sum\nolimits_{e\in\hat{H_j}} \left(1 - 2\cdot \mathbf{1}_{j+1\le j_t(e)}\right)\cdot g(\hat{f_e}), \tag{6}$$

where $\mathbf{1}_{j+1\le j_t(e)}$ is an indicator function that returns 1 if the element e is sampled for the layer $j+1$, and 0 otherwise.

4.2 Preliminary about KP-CF

Cuckoo Filter along Kicking Path (KP-CF) consists of an array of buckets, each of which contains s slots and forms a bucket-level min-heap. The bucket-level min-heap is used to record the IDs and frequencies of heavy hitters (HHs), and ensures that the minimum frequency element within the min-heap is maintained at the root node. We give an example in Fig. 3 to explain its design.

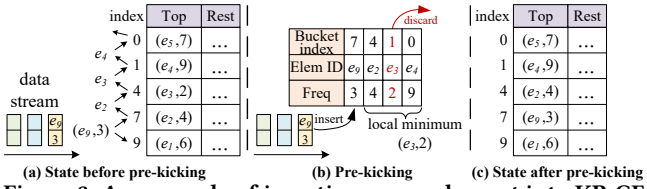


Figure 3: An example of inserting a new element into KP-CF

State before Pre-Kicking. In Fig. 3(a), we only show the element at the root node of the bucket-level min-heap in the "Top" column to ease the presentation. For each arrival data stream tuple, it is hashed to two candidate buckets by Cuckoo hash [77]. For example, the tuple $(e_9, 3)$ is hashed to buckets 7 and 9. If both of the candidate buckets are full, the KP-CF chooses bucket 7 to kick, which contains the record $(e_2, 4)$ with the minimum frequency 4 in the two candidate buckets. So, as is illustrated in Fig. 3(a), the record $(e_2, 4)$ in bucket 7 is to be kicked to bucket 4, and the record $(e_3, 2)$ in bucket 4 is to be kicked to bucket 1, and so on. The procedure continues until an empty slot is found or a predefined number of kicks, which we call *Maxkicks*, is reached. In Fig. 3(a), we assume the *Maxkicks* is 4 and all the 4 kicks encounter full buckets. At the end of the procedure, we can select the last kicked element, the element e_5 whose frequency is 7 in Fig. 3(a), as the victim element to be discarded from the KP-CF. However, the element e_3 has the lowest frequency among the set of all the kicked elements $\{e_2, e_3, e_4, e_5\}$. If we choose e_3 as the victim, we can achieve the best maintenance of the HHs under the constraints of the *MaxKicks*.

Pre-Kicking. This procedure aims to identify the aforementioned victim element, namely, the kicked element with the lowest frequency on the kick-out path of a cuckoo hash table. So KP-CF uses a pre-kicking queue in Fig. 3(b) to record the information of kicked elements in each round in the form of 3-tuples. For example, the arrival tuple $(e_9, 3)$, which is to be inserted into the bucket 7, is recorded by $(7, e_9, 3)$. The record $(e_2, 4)$, which is currently in bucket 7 and is to be kicked to the bucket 4, is recorded by $(4, e_2, 4)$. After that, KP-CF chooses the local minimum element in the pre-kicking queue as the victim to discard, which is e_3 in Fig. 3(b).

State after Pre-Kicking. In Fig. 3(c), according to the pre-kicking queue, KP-CF inserts $(e_9, 3)$ into the top slot of bucket 7, $(e_2, 4)$ into the top slot of bucket 4, and adjusts the min-heaps of these buckets.

5 REMOVABLE UNIVERSAL SKETCH

We first describe the structure and workflow of Removable Universal Sketch (RUS), followed by an introduction to its applications.

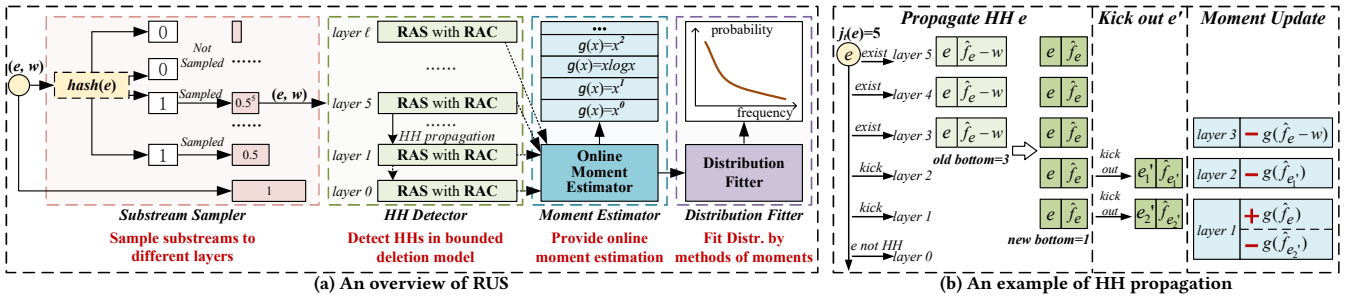
5.1 Overview of RUS

As depicted in Fig. 4(a), RUS is facilitated by four sequentially interconnected modules. Each module in Fig. 4(a) processes the output from its predecessor, beginning with the *Substream Sampler*, which receives a tuple (e, w) consisting of an element e and a weight w from the bounded deletion stream. The tuple represents an insertion event when w is positive and a deletion event when w is negative.

Substream Sampler. As shown in Fig. 4(a), the *HH Detector* in RUS is constructed as a layered structure, and we employ the *Substream Sampler* to sample arrival stream tuples to different layers, forming substreams for each layer. Specifically, for each arrival tuple (e, w), the sampler uses Eq. (4) to calculate the topmost sampled layer of the element e (e.g., layer 5 in Fig. 4), and delivers the tuple (e, w) to the *HH Detector* at this layer, indicating that e is sampled to the layers from the 0th layer to the topmost sampled layer.

HH Detector. As shown in Fig. 4(a), there are $\ell + 1$ layers in the *HH Detector* module, where each layer is used to identify the heavy hitters from the substream sampled to this layer. For example, in Fig. 4(a), the topmost sampled layer of the element e is layer 5, so the element e is in the substreams sampled for layer 0 to layer 5, and the *HH Detector* has to identify whether the element e is a heavy hitter within the substreams sampled for these layers. To achieve this, firstly, the 5th-layer of *HH Detector* identifies whether the element e is a heavy hitter at this layer. Then, if yes, the *HH Detector* propagates the element e with its frequency $\hat{f_e}$ estimated from layer 5 to lower layers, to see if the element e is also a heavy hitter at these layers. We detail the two procedures as follows.

1) HH Identification. We propose the Removable Augmented Sketch (RAS), detailed in **Section 6**, which functions as a layer within the multi-layered *HH Detector* to identify HHs from a substream sampled from the *Substream Sampler* module. In our design, RAS comprises an *element-frequency table* and a CountSketch (CS), as shown in Fig. 2(b). For the *element-frequency table*, we design it based on the KP-CF to improve query efficiency. For the CS, we substitute the conventional 32-bit integer counter with our proposed



(a) An overview of RUS (b) An example of HH propagation

Figure 4: An overview of RUS and an example of HH propagation

16-bit active counter, namely the Removable Active Counter (RAC), to double the number of allocated counters and thereby enhance the frequency estimation accuracy [76, 77]. In addition to the common addition operation, RAC supports subtraction operation to handle deletions in the *bounded deletion model* (see **Section 7** for details).

2) HH Propagation. Fig. 4(b) shows an example of *HH propagation*. We assume that, the element e has been identified as a heavy hitter at layers 3 to 5, with frequency estimation $\hat{f}_e - w$. Consider an arrival tuple (e, w), where w is positive, and $j_t(e) = 5$ denotes the topmost sampled layer of element e. We first update the frequency of the element e recorded at layer 5 from $\hat{f}_e - w$ to $\hat{f}_e$. Then, we propagate $(e, \hat{f}_e)$ to lower layers. For layers 4 and 3, we directly update the frequency of e from $\hat{f}_e - w$ to $\hat{f}_e$ as e is already recorded as a heavy hitter at these layers. For lower layers, if the *element-frequency table* used to record heavy hitters is full, we check if $\hat{f}_e$ is greater than any of the heavy hitters at these layers. Subsequently, we find the frequency of the elements e_1' and e_2' is lower than $\hat{f}_e$ at layers 2 and 1, respectively. So we kick out the two elements from the *table* to make room for e. For layer 0, we do nothing since $\hat{f}_e$ is not larger than any of the frequencies of recorded heavy hitters.

Online Moment Estimator. After obtaining HHs from each of the $\ell + 1$ layers of the *HH Detector*, the *recursive summation technique* shown in Eq. (6) can be used to estimate generic moments. However, as introduced in Section 4.1, this technique is limited to offline mode due to its requirement for a recursive scan of all HHs from the ℓth-layer down to the 0th-layer. So we propose an Online Moment Estimator (OME) in **Section 8** to update the moment estimation as soon as the set of HHs changes, i.e., the *HH Detector* finds a new HH, or finds that a HH element no longer qualifies as a HH.

Distribution Fitter [89]. The frequency moments can be utilized to reconstruct the frequency distribution in real time by employing the distribution fitter based on the method of moments and the cut-then-rejoin strategy. For further details, please refer to [89].

5.2 Applications of RUS

Network Measurement. Network measurement utilizes only the size-limited SRAM (usually in MBs) on their line cards to measure multiple kinds of statistics, including per-packet frequency, heavy hitters, network flow cardinality, network flow frequency distribution and its entropy [72]. In addition, network administrators may delete the network flow records relating to network attacks, which only account for a small fraction of all the records collected. So, our RUS can be used to monitor and manage IP network services [24].

Database Query Optimization. Query optimization aims to identify an execution plan for a query statement that minimizes intermediate results. To achieve this, DBMS has to collect basic statistics for each column, such as per-value frequency (i.e., per-element frequency), the number of distinct values (i.e., the 0th-order moment), the top frequency histogram (i.e., HHs), and quantile information (i.e., frequency distribution), while managing record insertions and partial deletions in high speed [28]. Considering the database's size, there are limitations on the memory resources to retrieve these statistics, e.g., memory in MBs in the Join Order Benchmark [18]. So our methods show promise for application in query optimization.

N-Gram Mining. N-gram mining is extensively employed in natural language processing applications [40, 44, 62], such as sequence embedding [43, 51] and sequence classification [57]. A significant challenge in n-gram mining is the combinatorial explosion of token combinations. For instance, a corpus with only 300 distinct words can yield up to 90,000 distinct bigrams. Consequently, the memory footprint required to exactly record the statistics of these n-grams could be substantial. Additionally, the enforcement of data protection laws and the right to be forgotten [41, 53, 65] requires the removal of specific information from the corpus in the trained models or collected statistics, highlighting the necessity for deletion support. The function of our RUS meets the needs of this scenario.

6 REMOVABLE AUGMENTED SKETCH

In this section, we propose a Removable Augmented Sketch (RAS) to track heavy hitters (HHs) under the *bounded deletion model*.

6.1 Design Rationale

For the problem of detecting ε-heavy hitters in the bounded deletion model, a naïve solution is to use a min-heap to hold the top-k heavy hitters. However, this method is impractical due to its low element processing throughput. Whenever an element arrives, we need to search the element ID in the min-heap, which involves a linear scan with $O(k)$ time cost. The number k of HHs that need to be held in the min-heap, under the bounded deletion model, can be very large, e.g., $k = 8192$ when $\zeta = 12$ and $\varepsilon = 2^{-13}$, as shown in Section 6.4.

An improved solution is to additionally maintain a hashtable to reduce the key lookup time complexity to $O(1)$, like SpaceSaving$^{\pm}$ (SS$^{\pm}$) [48, 90]. However, in addition to the extra memory cost for the hashtable, the heap maintenance is still time-consuming. Essentially, the min-heap is a sorting data structure for quickly locating the *global minimum element* at the root. When updating the frequency of an existing element or inserting a new element, it needs to restore the min-heap property, involving $O(\log k)$ sift up/down operations.

Thus, we embrace the design of the KP-CF [77], which searches only the local minimum along the kicking path and uses it as the victim to make room for the new element. KP-CF is memory efficient as its load factor can be up to 95% [21], and it has both $O(1)$ lookup and insertion time costs, the same as the Cuckoo Filter. However, its shortcoming is to kick only the path-level local minimum element for replacement. We address this by using the KP-CF as the prefilter and using a backend sketch to store the kicked victim elements.

6.2 Overview of RAS

As shown in Fig. 5, our RAS has two main components: a prefilter and a sketch. The prefilter consists of two parts: a Cuckoo Filter along Kicking Path (KP-CF) and a tiny histogram. The KP-CF serves as the function of the table in Fig. 2(b), which keeps the IDs and frequencies of the HHs. The histogram is used to track the minimum frequency of the HHs, since KP-CF cannot be queried for the minimum frequency of the HHs as the min-heap does.

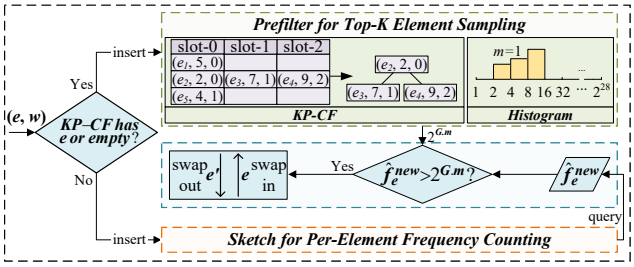


Figure 5: Structure and workflow of our RAS

We explain the necessity of the histogram through an example. Suppose we use the KP-CF in Fig. 3 as a prefilter to track HHs. If an insertion event $(e_{10}, 1)$ arrives carrying an untracked element and a weight lower than the frequencies of all the tracked elements, the element e_{10} is not going to be recorded by the KP-CF. However, the pre-kicking will be triggered blindly as the KP-CF does not maintain the global minimum frequency of the HHs. This blind attempt will cause significant computational waste, as most elements (e.g., e_{10}) are not HHs. Thus, we use a tiny histogram to track the *approximate global minimum frequency* of the HHs. We insert an element into the KP-CF only if its frequency exceeds this *minimum frequency*.

KP-CF Overview. As shown in Fig. 5, the KP-CF we use is almost the same as the version shown in Fig. 3, except that each slot maintains three attributes of an element: (id, newFreq, oldFreq). The newFreq of an element denotes its current frequency. The oldFreq of an element is the frequency accumulated during its residence in the sketch before being inserted into the KP-CF, which is 0 if the element has never resided in the sketch. The difference between oldFreq and newFreq of an element is the exact aggregated frequency that has been accumulated during its residence in the KP-CF [55].

Histogram Overview. As shown in Fig. 5, the histogram is an array, and its ith entry keeps the number of HHs whose frequency falls in the range of $[2^i, 2^{i+1})$, $0 \le i \le 27$. We set up an indicator m to record the index of the first non-zero entry. So 2^m represents the *approximate global minimum* (*approx-min* for short) frequency of the elements in the KP-CF. The histogram has three operations:

- **Insert**: If a new element e with frequency f is inserted into the KP-CF, we calculate $i = \lfloor \log_2(f) \rfloor$ and increase the ith entry of the histogram by 1. Then, we assign i to m if i is less than m.
- **Remove**: If an old element e with frequency f is removed from the KP-CF, we decrease the ith entry of the histogram by 1, where $i = \lfloor \log_2(f) \rfloor$. Then, if m equals i and the ith entry becomes 0, we update m to the index of the next non-zero entry.
- **Update**: If an existing element e in the KP-CF has its frequency updated, this operation can be seen as one remove and one insert.

With this histogram, we can easily decide whether to insert an element into the KP-CF by comparing its frequency with 2^m.

6.3 Algorithm for Updating RAS

We present the update procedure of RAS in Algorithm 1. At Line 1, we use two regular hash functions, $h_1(\cdot)$ and $h_2(\cdot)$, to map elements to candidate buckets. We then describe the algorithm in four phases.

Algorithm 1: Update procedure of RAS

Input: an arrival tuple consisting of element ID e and weight w
State: prefilter R with KP-CF B and histogram G, sketch M

1 $h_1(e) = hash(e)$, $h_2(e) = h_1(e) \oplus hash(e, salt)$
2 $\langle u, v\rangle$ = `lookupFilter`(R, e), $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle = \langle$NIL, 0, 0$\rangle$
3 **if** $u \neq$ NIL $\wedge\ v \neq$ NIL $\wedge\ R.B[u][v].id == e$ **then** // e exists
4 `updateFilter`$(R, u, v, \hat{f}_e^{new} = w + R.B[u][v].newFreq)$
5 **else if** $u \neq$ NIL $\wedge\ v \neq$ NIL **then**// an empty slot exists
6 $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle$ = `insertFilter`$(R, e, \hat{f}_e^{new} = w, \hat{f}_e^{old} = 0)$
7 **else if** $(\hat{f}_e^{new}$ = `updateSketch`$(M, e, w)) > 2^{R.G.m}$ **then**
// there is no e or empty slot, but e's freq. $\hat{f}_e^{new} >$ minimum
8 $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle$ = `insertFilter`$(R, e, \hat{f}_e^{new}, \hat{f}_e^{old} = \hat{f}_e^{new})$
9 **if** $e' \neq$ NIL **then** `updateSketch`$(M, e', \hat{f}_{e'}^{new} - \hat{f}_{e'}^{old})$ // kick e'
10 **Function** `lookupFilter`(R, e): //prefilter, element
11 **foreach** $u \in \{h_1(e), h_2(e)\}, v \in [0, s)$ **do**
12 **if** $R.B[u][v].id == e$ **then return** $\langle u, v\rangle$
13 **if** $R.B[h_1(e)][0].id ==$ NIL **then return** $\langle h_1(e), 0\rangle$
14 **if** $R.B[h_2(e)][0].id ==$ NIL **then return** $\langle h_2(e), 0\rangle$
15 **return** $\langle$NIL, NIL$\rangle$
16 **Function** `updateFilter`$(R, u, v, \hat{f}_e^{new})$: // bkt idx, slot idx, freq.
17 update the histogram $R.G$ and $R.B[u][v].newFreq = \hat{f}_e^{new}$
18 restore the min-heap property of the bucket $R.B[u]$
19 **Function** `insertFilter`$(R, e, \hat{f}_e^{new}, \hat{f}_e^{old})$: //id, new&old freq.
20 initialize $R.Q$ with $R.Q[0]$ recording $\langle e, \hat{f}_e^{new}, \hat{f}_e^{old}\rangle$
21 fill pre-kicking queue $R.Q$ until an empty slot is recorded by $R.Q$ or the number of attempts reaches $Maxkicks$
22 **if** *no empty slot is recorded by* $R.Q$ **then**
23 find the minimum record $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle$ in queue $R.Q$
24 **if** $e' == e$ **then return** $\langle e, \hat{f}_e^{new}, \hat{f}_e^{old}\rangle$ // insert failed
25 update the histogram $R.G$ and update KP-CF $R.B$ by $R.Q$
26 **if** *an empty slot is recorded by* $R.Q$ **then return** $\langle$NIL, 0, 0$\rangle$
27 **else return** $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle$ // old element replaced by e

Lookup e. When a stream tuple (e, w) arrives, at Line 2, we call `lookupFilter`(R, e) to check whether the element e is present in the KP-CF $R.B$ or if any of e's candidate buckets have an empty slot.

Update when e Exists. At Line 3, if the index u and v are not NIL, and the id in the slot $R.B[u][v]$ is e, we confirm that e is present

in the slot $R.B[u][v]$. Consequently, to update e's newFreq, we call updateFilter($R, u, v, \hat{f}_e^{new}$=w+$R.B[u][v].newFreq$) at Line 4. Finally, the procedure ends as the condition at Line 9 does not hold.

Update when an Empty Slot is Found. If Line 5 holds, e's candidate buckets have at least one empty slot. So we insert the element e into the empty slot by insertFilter($R, e, \hat{f}_e^{new}$=$w, \hat{f}_e^{old}$=0). At Line 20, we initialize the pre-kicking queue $R.Q$. Then, as we have confirmed that an empty slot is in one of e's candidate buckets, no more elements will be inserted into the queue $R.Q$ at Line 21. Thus, at Line 25, the element e is inserted into an empty slot, and the histogram $R.G$ is updated accordingly. Next, as the condition at Line 26 holds, we return $\langle$**NIL**, 0, 0$\rangle$ to indicate that the insertion succeeds and no element is kicked out. Finally, the procedure ends.

Update after Querying Sketch. In this case, at Line 7, we first insert the tuple (e, w) into the sketch M and query for e's current frequency $\hat{f}_e^{new}$ from it. Then, if $\hat{f}_e^{new}$ does not exceed the *approx-min* $2^{R.G.m}$ recorded by the histogram $R.G$, the procedure ends as the frequency of item e is not greater than that of any item in the KP-CF. Otherwise, we try to insert the element e into the KP-CF $R.B$ by insertFilter($R, e, \hat{f}_e^{new}, \hat{f}_e^{old}$=$\hat{f}_e^{new}$) at Line 8. At Line 21, after initializing the pre-kicking queue at Line 20, we fill the queue until the number of kicks exceeds $Maxkicks$ or an empty slot is recorded.

If no empty slot is recorded at Line 21, the condition at Line 22 holds, and the limit $Maxkicks$ is reached. At Line 23, we find the element e' with the minimum frequency $\hat{f}_{e'}^{new}$ in the queue $R.Q$. If the minimum frequency element e' is e, we return $\langle e, \hat{f}_e^{new}, \hat{f}_e^{old}\rangle$ at Line 24 to indicate that the insertion fails. Otherwise, if e' is not e, the element e can be inserted into the KP-CF $R.B$, so we update the histogram $R.G$ at Line 25 and update the KP-CF $R.B$ by the pre-kicking queue $R.Q$ as illustrated in Fig. 3(c). Finally, at Line 27, we return the information $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle$ on the kicked element e', in order to swap out the element e' to the sketch M at Line 9.

If an empty slot is recorded at Line 21, the condition at Line 22 does not hold. At Line 25, we update the histogram $R.G$ and update the KP-CF $R.B$ by $R.Q$, and return $\langle$**NIL**, 0, 0$\rangle$ at Line 26 to indicate that the insertion succeeds and no element is kicked out. Then, the procedure ends as the condition at Line 9 does not hold.

6.4 Analysis of Size Lower Bound for Prefilter

Theorem 1 shows the lower bound of the prefilter size k to detect ε-HHs in the bounded deletion model. Its formal proof is omitted in this paper and can be found in our technical report [1]. Here, we provide only the outline of the proof. Let η be the skewness of element frequency distribution. We have $\eta \geq 1$ in most scenarios [31, 55], and $\eta = 1$ indicates low skewness. Under this assumption, we proved in [1] that the prefilter size can be set to $\frac{\zeta}{12\varepsilon}$. The error bounds for HH detection and frequency estimation are also given.

Theorem 1 (Prefilter capacity lower bound). *Assume that a bounded deletion stream has N distinct elements with $D : I$ ratio upper bounded by $1 - \frac{1}{\zeta}$, and has its per-element frequency following a Zipf distribution with skewness η. To detect the ε-HHs in this data stream, the capacity k of the prefilter must be at least $\left\lfloor \sqrt[\eta]{\zeta/\varepsilon \sum_{i=1}^{N} (1/i)^{\eta}} \right\rfloor$.*

Proof outline. In the bounded deletion stream, after all I insertions and D deletions, the frequency of each ε-HH is at least $\frac{\varepsilon}{\zeta} I$, given that $\varepsilon(I - D) = \frac{\varepsilon}{\zeta} I$. Thus, the prefilter must retain all elements with a frequency exceeding $\frac{\varepsilon}{\zeta} I$ before deletions occur. Otherwise, if such an element is not retained before deletions occur, it may become an untracked HH afterward. Therefore, the lower bound is the maximum value of k satisfying $f(k; \eta, N) \geq \frac{\varepsilon}{\zeta}$, where $f(i; \eta, N) = 1/\left(i^{\eta} \sum_{j=1}^{N} j^{-\eta}\right)$ is the probability mass function of Zipf distribution. By solving $f(k; \eta, N) \geq \frac{\varepsilon}{\zeta}$ for k, we have Theorem 1.

7 REMOVABLE ACTIVE COUNTER

We present the Removable Active Counter (RAC), which supports not only the common addition operation but also the subtraction operation to accommodate deletions in the bounded deletion model. We first show the internal representation of RAC, followed by an example to explain its subtraction operation. Finally, we discuss an optimization to improve the processing speed when using RAC. For the pseudocode of the addition and subtraction operations, as well as mathematical analysis, please refer to our technical report [1].

7.1 Internal Representation of RAC

An RAC is a compressed 16-bit representation of a signed integer. The internal structure of an RAC C consists of three parts:

- The **sign bit** $C.\boldsymbol{\rho}$ is the leading bit;
- The **exponent part** $C.\boldsymbol{\alpha}$ occupies the subsequent $\mathcal{L}_\alpha = 4$ bits;
- The **coefficient part** $C.\boldsymbol{\beta}$ contains the remaining $\mathcal{L}_\beta = 11$ bits.

To further save memory, for the coefficient $C.\beta$, we use a leading-one convention, which assumes that its leftmost bit is always 1 and is not stored explicitly. We denote the coefficient under the leading-one convention as $\tilde{\beta} = 2^{\mathcal{L}_\beta} + C.\beta$, where $\mathcal{L}_\beta$ is the length of the coefficient part, defaulting to 11. To find the represented integer value of an RAC C, the evaluation function V can be used:

$$V(C) = (2\,C.\rho - 1) \cdot \left((2^{\mathcal{L}_\beta} + C.\beta) \cdot 2^{C.\alpha} - 2^{\mathcal{L}_\beta}\right), \tag{7}$$

where $(2\,C.\rho - 1)$ converts the sign bit $C.\rho$ to ± 1. The initial values of $C.\alpha$, $C.\beta$ and $\tilde{\beta}$ are 0, 0 and $2^{\mathcal{L}_\beta}$, respectively, since $\tilde{\beta}$ carries an implicit leftmost 1 bit ahead of $C.\beta$. The last term $-2^{\mathcal{L}_\beta}$ provides a bias correction to ensure the initial value of $V(C)$ is zero. The maximum values of $C.\alpha$ and $\tilde{\beta}$ are $2^4 - 1 = 15$ and $2^{12} - 1 = 4095$, respectively. Thus, the counting range of our RAC is within $\pm(4095 \cdot 2^{15} - 2^{11})$, which is roughly $\pm 2^{27}$, sufficient for most practical applications.

7.2 Example of Subtraction Operation in RAC

Fig. 6 shows an example to subtract $|w|$ from an RAC C. Assume an RAC C with $\rho = 1b$ (denoted by +), $\alpha = 001b$, $\tilde{\beta} = ①001b$, and $\mathcal{L}_\beta = 3$, where ① represents the implicit leftmost 1 bit. The weight w in our example is $-16 = -1 \cdot 2^{001b} \cdot 1000b$. Since the counter value $V(C)$ has been multiplied by $2^{C.\alpha}$ in Eq. (7), we need to divide the increment w by $2^{C.\alpha}$, so that it can be added to the coefficient $\tilde{\beta}$. We denote the division result as $\Delta\tilde{\beta}$, yielding $-1 \cdot 1000b$ in our example in Fig. 6. We show the example in three phases.

- **Obtain new $\tilde{\boldsymbol{\beta}}$.** As shown in Fig. 6, we obtain the new value of $\tilde{\beta}$ by summing $\Delta\tilde{\beta}$ with $\tilde{\beta}$, resulting in $-1 \cdot 1000b + ①001b = ⓪001b$.
- **Update $\boldsymbol{\rho}$ and $\tilde{\boldsymbol{\beta}}$.** After the update of $\tilde{\beta}$, the value of $V(C)$ becomes negative as $\tilde{\beta} \cdot 2^{C.\alpha} - 2^{\mathcal{L}_\beta}$ falls below 0. However, $2C.\rho - 1$ is still +1, contradicting the representation of RAC, which assumes

that the term $\tilde{\beta} \cdot 2^{C.\alpha} - 2^{\mathcal{L}_\beta}$ is non-negative, and the sign of $V(C)$ matches $2C.\rho - 1$. So, in Fig. 6, we flip the sign bit $C.\rho$ from $1b$ to $0b$ to ensure $2C.\rho - 1$ equals -1. For the term $\tilde{\beta} \cdot 2^{C.\alpha} - 2^{\mathcal{L}_\beta}$, we recalculate the value of $\tilde{\beta}$ according to Eq. (8), so that the term adheres to the representation in Eq. (7) while yielding a value equal to $-(\tilde{\beta} \cdot 2^{C.\alpha} - 2^{\mathcal{L}_\beta})$. So, in Fig. 6, we assign ⓪$111b$ to $\tilde{\beta}$.

$$-(\tilde{\beta} \cdot 2^{C.\alpha} - 2^{\mathcal{L}_\beta}) \rightarrow \underbrace{(2^{\mathcal{L}_\beta - \alpha + 1} - \tilde{\beta})}_{new\ value\ of\ \tilde{\beta}} \cdot 2^{\alpha} - 2^{\mathcal{L}_\beta} \tag{8}$$

• **Correct underflow**. The resulting $\tilde{\beta}$ experiences underflow because its left-most implicit bit is ⓪. To correct it, we adjust the value of $\tilde{\beta}$ and α to ensure the implicit bit of $\tilde{\beta}$ becomes ①. As illustrated on the right-hand side of Fig. 6, when $\tilde{\beta}$ becomes ⓪$111b$, we left-shift $\tilde{\beta}$ by 1 bit, and decrement α by 1 accordingly.

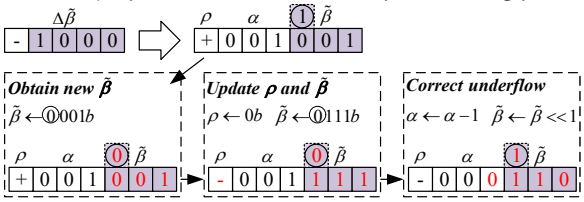


Figure 6: Example of updating an RAC C

7.3 Optimization

With RAC, the estimation accuracy of traditional sketches, such as CountSketch (CS) and CountMin Sketch, can be improved. However, for each insertion or deletion, these sketches update one counter per row, resulting in the update of d (e.g., 4, 8) RACs. This leads to a higher processing overhead compared to using traditional integer counters, due to the complexity of the RAC's update procedure.

To mitigate this, a good practice is to limit updates to only one counter per insertion or deletion, as demonstrated by Virtual Active Counter [47, 94]. Thus, we introduce a variant of CS, Single-update CountSketch (SuCS), which diverges from CS in three primary aspects: first, SuCS maintains a one-dimensional array; second, it uses d hashes to find d counters, but randomly updates only one of them; third, it returns 1 if the query result of frequency is less than 0, to align with the bounded deletion model. In RAS, we replace CountSketch with SuCS, and then apply RAC to SuCS, yielding RA-SuCS. Our experiments show that, RA-SuCS provides an effective trade-off between estimation accuracy and processing speed. The mathematical analysis is available in our technical report [1].

8 ONLINE MOMENT ESTIMATOR

We first show the rationale behind the Online Moment Estimator (OME), followed by an example of online frequency moment estimation. The pseudocode is available in our technical report [1].

8.1 Rationale behind Online Moment Estimator

The *recursive summation technique* [7] estimates the frequency moment of all layers recursively from layer ℓ to layer 0 by Eq. (6). We present the following general formula to estimate the 0th-layer moment $\hat{L_0}$ without recursion. This is achieved by applying the ℓth-layer moment $\hat{L_\ell} = \sum_{e \in \hat{H}_\ell} g(\hat{f}_e)$ to its lower-layer moment $\hat{L}_{\ell-1}$ in Eq. (6), and then recursively to $\hat{L}_{\ell-2}$, and so on.

$$\hat{L_0} = 2^\ell \sum\nolimits_{e \in \hat{H}_\ell} g(\hat{f}_e) + \sum\nolimits_{j=0}^{\ell-1} 2^j \sum\nolimits_{e \in \hat{H}_j} \left(1 - 2 \cdot \mathbf{1}_{j+1 \le j_t(e)}\right) g(\hat{f}_e)$$

According to $j_t(e)$ defined in Eq. (4), an element e is never sampled on layer $\ell + 1$, i.e., $\mathbf{1}_{\ell+1 \le j_t(e)} \equiv 0$. So we rewrite this equation to

$$\hat{L_0} = \sum\nolimits_{j=0}^{\ell} 2^j \sum\nolimits_{e \in \hat{H}_j} \left(1 - 2 \cdot \mathbf{1}_{j+1 \le j_t(e)}\right) g(\hat{f}_e). \tag{9}$$

It implies that the moment can be estimated by summing up the moment contribution of each element on each layer, e.g., the contribution of the element e on the jth layer is $2^j \cdot \left(1 - 2 \cdot \mathbf{1}_{j+1 \le j_t(e)}\right) \cdot g(\hat{f}_e)$.

Then, consider a case when all the elements are present in multiple adjacent layers. For example, an element e is present in each layer from $j_t(e)$ to $j_b(e)$, where $j_t(e)$ is e's topmost sampled layer, and $j_b(e)$ is the lowest layer among all the layers that regard e as a heavy hitter. We call $j_b(e)$ the *bottom layer* of the element e:

$$j_b(e) = \min j, \quad \text{subject to } 0 \le j \le \ell \ \wedge \ e \in \hat{H}_j. \tag{10}$$

In this case, Eq. (9) can be converted to

$$\begin{aligned} \hat{L_0} &= \sum\nolimits_{e \in \cup \hat{H}_j} \sum\nolimits_{j=j_b(e)}^{j_t(e)} 2^j \left(1 - 2 \cdot \mathbf{1}_{j+1 \le j_t(e)}\right) g(\hat{f}_e), \qquad (11) \\ &= \sum\nolimits_{e \in \cup \hat{H}_j} \left(2^{j_t(e)} - \sum\nolimits_{j=j_b(e)}^{j_t(e)-1} 2^j\right) g(\hat{f}_e) = \sum\nolimits_{e \in \cup \hat{H}_j} 2^{j_b(e)} g(\hat{f}_e) \end{aligned}$$

where $\cup \hat{H}_j$ is the union of all sets of heavy hitters from layer ℓ to 0.

So, we obtain OME as expressed by Eqs. (9) and (11). For an arrival tuple (e, w) carrying an element e that appears in multiple adjacent layers (e.g., *HH propagation*), we can incrementally update the moment $\hat{L_0}$ by Eq. (11), using the old values of $\hat{f}_e$ and $j_b(e)$ before the tuple's arrival and the new values afterward. In addition, if an element e' is no longer a heavy hitter on the jth layer (e.g., e' is kicked out during the *HH propagation*), we update the moment $\hat{L_0}$ by removing the contribution of e' on layer j using Eq. (9).

8.2 Example of Online Moment Updating

Combining the *HH Propagation* described in Section 5.1, we illustrate an example of the online moment updating in Fig. 4(b). Suppose an RUS with layers indexed from 0 to ℓ, and an arrival tuple (e, w), whose topmost sampled layer $j_t(e)$ is layer 5.

Update Moment by Arrival Element e. We show it in two phases.

• **Remove old contribution.** As shown on the left-hand side of Fig. 4(b), the element e is delivered to layer 5 and then propagates to layer 1. Before the arrival of the tuple (e, w), the bottom layer of the element e was layer 3, and it becomes layer 1 afterward. Thus, in the column "Moment Update" of Fig. 4(b), we remove the contribution of element e at layer 3 using Eq. (11), which is $2^{j_b^{old}(e)} g(\hat{f}_e - w)$, where $j_b^{old}(e)$ denotes e's old bottom layer, namely layer 3, and $\hat{f}_e - w$ denotes e's frequency before the arrival of the tuple (e, w).

• **Add new contribution.** As shown on the left-hand side of Fig. 4(b), layer 1 is the new bottom layer of e after propagation. So, as depicted in the "Moment Update" column of Fig. 4(b), we add the contribution of the element e at layer 1 using Eq. (11), calculated as $2^{j_b^{new}(e)} g(\hat{f}_e)$, where $j_b^{new}(e)$ represents e's new bottom layer, namely layer 1, and $\hat{f}_e$ denotes the current frequency of e.

Update Moment by Kicked Elements e_1', e_2'. During the propagation, an element e_1' is kicked out at layer 2. Consequently, as illustrated in the "Moment Update" column of Fig. 4(b), we remove the contribution of element e_1' at layer 2 using Eq. (9), which is $2^j \cdot \left(1 - 2 \cdot \mathbf{1}_{j+1 \le j_t(e_1')}\right) \cdot g(\hat{f}_{e_1'})$, where j is the current layer, i.e., 2, and

$j_t(e'_1)$ denotes the topmost sampled layer of e'_1. Similarly, we remove the moment contribution of the element e'_2 at layer 1 by Eq. (9).

9 EXPERIMENTAL EVALUATION

This section evaluates the performance of our proposed solutions for estimating per-element frequency, heavy hitters, and frequency moments in the *bounded deletion model*. Due to the page limit, we omit the experiments on parameter setting and frequency distribution estimation, which are detailed in our technical report [1].

9.1 Experimental Setup

Computation Platform. We conduct all the experiments on a 12-core CPU server (Intel i7-12700) with 128GB of memory.

Implementation. We implement the Single-update CountSketch (SuCS), CountSketch [10] with RAC (RA-CS), SuCS with RAC (RA-SuCS), Removable Augmented Sketch (RAS), Removable Universal Sketch (RUS), and all competing methods in different tasks in C++.

- **Per-element frequency estimation:** CMS [16], CS [10], randomized error-reduction Sketch (rSkt) [67], and Diamond Sketch (DS) [84]. We compare them with RA-SuCS, RA-CS, and SuCS. For DS, we set all parameters as suggested in the original paper [84]. We set the number of rows to 1 for SuCS and 4 for other methods.
- **Heavy hitter detection:** CMS/CS/DS/rSkt+MH (sketch with min-heap postfilter), MH+CMS/CS/DS/rSkt (sketch with min-heap prefilter) based on the frequency estimation methods, as well as SpaceSaving$^{\pm}$ (SS$^{\pm}$) [90]. We compare these methods with RAS.
- **Moment estimation:** UnivMon [42] and Off-RUS (an offline version of RUS that estimates moments using an offline moment estimator). We compare UnivMon and Off-RUS with our RUS.

Datasets. We use three real-world datasets and one synthetic dataset.

1) CAIDA Traces [9]. The CAIDA traces are collected from the Equinix Chicago high-speed monitor. We use the trace with a monitoring interval of 60s. The trace contains about 30.1M packets, originating from 1.1M distinct IP flows identified by (srcIP, srcPort, dstIP, dstPort, Protocol). We extract the ID of the IP flow from each packet to form the data stream, so there are about 30.1M insertions, and the insertions and deletions are represented as (flow ID, ±1).

2) IMDB Dataset [36]. The dataset includes multiple relations used for the Join Order Benchmark [36]. We select a high-cardinality column *person_role_id*, which contains about 17.6M rows with 3.1M distinct values, from the largest table *cast_info*. We extract the value ID to construct the data stream, so there are about 17.6M insertions, and the insertions and deletions are in the form of (value ID, ±1). In database systems, the statistics collected from a column can be used to estimate the result size of all sub-plans of each query by a query optimizer, such as the Selinger query optimizer [58, 71].

3) Yelp Reviews Dataset [88]. The dataset includes around 7.0M business reviews in chronological order. We perform bigram text mining on each review, and identify about 22.0M distinct bigrams with a total frequency of 733.9M. The corresponding data stream is formed by the bigrams extracted from each of the ordered reviews. Overall, there are about 733.9M insertions, and the insertion and deletion events are represented as (bigram ID, ±1).

4) Synthetic Zipf Distribution Dataset [90]. The dataset contains about 4.9M distinct items with a total frequency of 20.0M, where items' frequencies follow Zipf's Law [45] with skewness $\eta = 1$. The corresponding data stream includes 20.0M insertions, and the insertion and deletion events are represented as (item ID, ±1). We explore two arrival patterns of the dataset to show the robustness of our solution. In the *shuffled pattern*, we shuffle the dataset so that items arrive in a random order. In the *sorted pattern*, we sort the dataset to allow the items with lower frequencies to arrive first.

Deletion Patterns. For the CAIDA traces, IMDB, and Zipf distribution datasets, deletions in the corresponding data streams are randomly chosen from the insertions. For the Yelp reviews dataset, deletions are made from the oldest reviews. Therefore, the insertion events that arrive first in the data stream are deleted first.

Metrics. We use the Average Absolute Error (AAE), Average Relative Error (ARE), F1 Score (F1), Relative Error (RE), and Throughput in Millions of Operations per Second (Mops) as our evaluation metrics. These metrics are detailed in our technical report [1].

9.2 Performance of Frequency Estimation

On the three real-world datasets, we compare RA-SuCS against RA-CS, SuCS, CMS, CS, DS, and rSkt under different memory allocations, with the $D : I$ ratio set to 0.5. We set the memory allocation ranges from 128KB to 1024KB for the CAIDA traces and IMDB datasets, and from 5MB to 50MB for the Yelp reviews dataset, as the number of distinct bigrams extracted significantly exceeds the count of unique elements in the first two datasets. The AAE and ARE for the three real-world datasets are shown in Figs. 7a–7f.

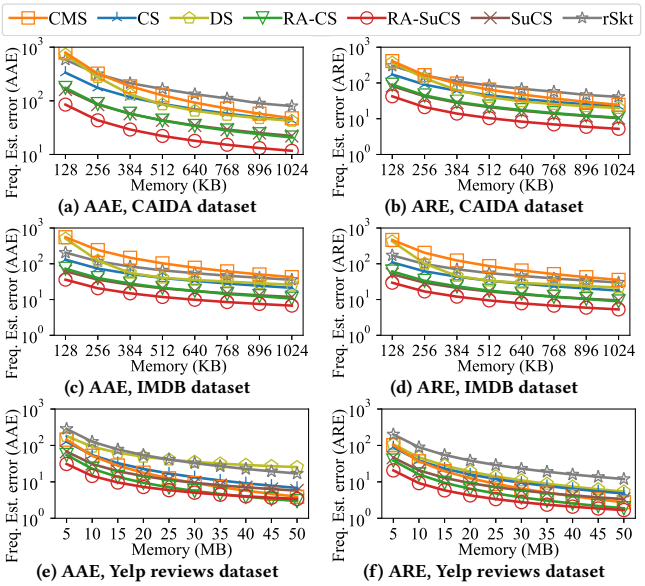


Figure 7: Frequency estimation errors in three datasets

AAE (Figs. 7a, 7c and 7e). Compared with CMS, CS, RA-CS, DS, SuCS, and rSkt, the AAE of RA-SuCS is 12.4, 3.7, 2.0, 11.3, 1.8, and 6.3 times lower, respectively, on average under 128KB of memory for the CAIDA traces and IMDB datasets; it is 4.6, 4.1, 1.8, 5.8, 2.2, and 9.2 times lower under 5MB of memory for the Yelp reviews dataset.

ARE (Figs. 7b, 7d and 7f). Compared with CMS, CS, RA-CS, DS, SuCS, and rSkt, the ARE of RA-SuCS is 12.7, 3.8, 2.0, 11.5, 1.8, and 6.5 times lower, respectively, on average under 128KB of memory for

the CAIDA traces and IMDB datasets; it is 4.9, 4.3, 1.9, 5.2, 2.2, and 9.7 times lower under 5MB of memory for the Yelp reviews dataset.

Analysis of Frequency Estimation. (a) SuCS outperforms other methods that use 32-bit integer counters, including CMS, CS, and rSkt, in our evaluation. This is because it uses the Tug-of-War technique like CS to reduce the cumulative estimation error, rather than adding up all frequencies into the hashed counter like CMS, which increases the cumulative error. Besides, SuCS sets the estimation to 1 if it falls below 0, which complies with the bounded deletion model where the deletion is bounded. (b) RA-SuCS shows the best performance, since it combines SuCS with 16-bit RAC, further enhancing the estimation accuracy by doubling the number of counters. (c) DS also proposes a counter compression technique, but it does not show a performance advantage. This is because it is designed for low-skew (e.g., skewness between 0 and 1) data streams carrying low-frequency elements (e.g., frequencies less than 2^{16}) [84].

9.3 Performance of Heavy Hitter Detection

We compare RAS with CMS/CS/DS/rSkt+MH, MH+CMS/CS/DS/rSkt, and SS$^{\pm}$ in HH detection. We conduct two experiments to show the performance under fixed and varying $D:I$ ratios, respectively. In both experiments, we set the threshold ε for HH detection as defined in Eq. (2) to 2^{-14}, and set the filter size to $\frac{\zeta}{12\varepsilon}$ by Section 6.4.

1) We fix the $D:I$ ratio (i.e., $1-\frac{1}{\zeta}$) to 0.5 and use the CAIDA traces dataset, IMDB dataset, and Yelp reviews dataset for evaluation. We set the memory allocation to range from 256KB to 1024KB.

2) With 256KB of memory, we use the synthetic Zipf distribution dataset to evaluate the performance of HH detection when the $D:I$ ratio ranges from 0.1 to 0.9, which reflects the bounded deletion scenario, where deletions exist and the $D:I$ ratio is bounded.

9.3.1 ***Performance under Fixed $D:I$ Ratio.*** The results for the experiments under a fixed $D:I$ ratio are shown in Figs. 8a–8f.

F1 Score of HH Detection (Figs. 8a, 8c and 8e). Compared with CMS+MH, CS+MH, DS+MH, MH+CMS, MH+CS, MH+DS, MH+rSkt, SS$^{\pm}$, and rSkt+MH, the F1 Score of our RAS is 21%, 53%, 55%, 20%, 16%, 18%, 35%, 59%, and 69% higher, respectively, on average under 256KB of memory for the three real-world datasets.

ARE of HH Frequency Estimation (Figs. 8b, 8d and 8f). Compared with CMS+MH, CS+MH, DS+MH, MH+CMS, MH+CS, MH+DS, MH+rSkt, SS$^{\pm}$, and rSkt+MH, the ARE of our RAS is 30.6, 32.2, 37.6, 23.4, 7.1, 6.3, 21.7, 28.3, and 39.5 times lower, respectively, on average under 256KB of memory for the three real-world datasets.

Analysis of HH Detection under Fixed $D:I$ Ratio. (a) The prefilter-based methods outperform both the postfilter- and counter-based methods. In postfilters, the frequencies recorded are solely queried from the sketch, whereas in prefilters, updates to elements can be precisely recorded once they are loaded. For the counter-based method SS$^{\pm}$, when the table is full, it has to replace the newly arrived element with one already recorded in the table, leading to information loss and resulting in poorer overall accuracy of HH detection compared to methods that combine a sketch with a prefilter. (b) RAS performs better than the conventional prefilter-based methods, because we adopt a KP-CF as a prefilter and use RA-SuCS as the backend sketch. The prefilter improves the frequency estimation accuracy, and RA-SuCS outperforms other backend sketches, thus showing superior performance compared to conventional methods.

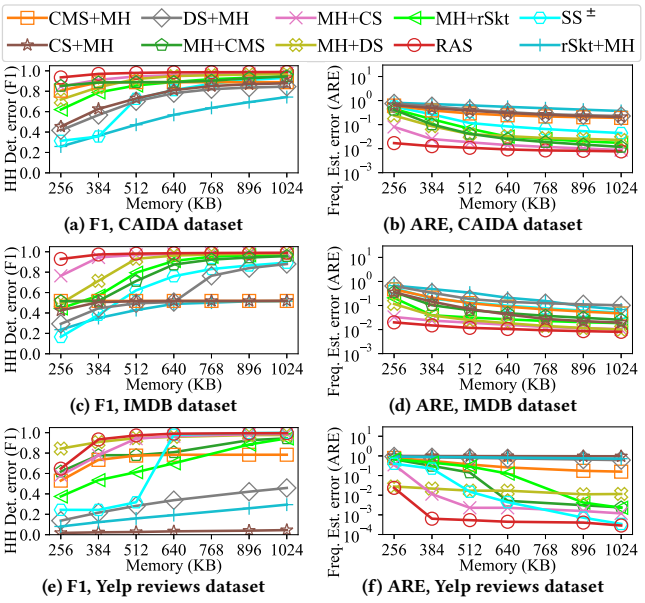


Figure 8: HH detection errors in three datasets

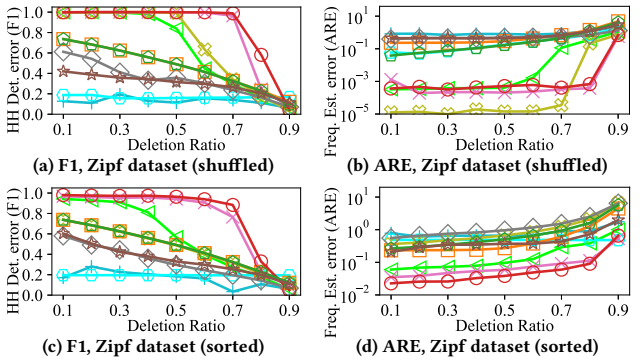


Figure 9: HH detection errors under varying deletion ratios

9.3.2 ***Performance under Varying D:I Ratios.*** The results for the shuffled pattern and sorted pattern of the Zipf distribution dataset described in Section 9.1 are shown in Figs. 9a–9d.

Analysis of HH Detection under Varying $D:I$ Ratios. (a) As shown in Fig. 9, as the $D:I$ ratio increases, the performance of the HH detection quantified by the F1 Score and ARE degrades accordingly. The reason is that, the lower bound of the prefilter size increases with the $D:I$ ratio, which reduces the memory budget for the backend sketch, thereby increasing the frequency estimation error and adversely affecting HH detection. (b) In the shuffled pattern, the frequency estimation accuracy of MH+DS is superior to that of other methods, whereas in the sorted pattern, it is not. The root cause is that, in the shuffled pattern, the high-frequency elements occur sooner than in the sorted pattern where the elements with low frequencies arrive first. Therefore, the true heavy hitters are more easily captured by the prefilter, which reduces the skewness of the data streams fed into the sketch, where DS excels [84]. However, in the sorted pattern, the heavy hitter elements must be stored in the sketch initially, leading to poorer performance of DS.

9.4 Performance of Moment Estimation

We compare RUS with UnivMon and Off-RUS in the 0th-, 2nd-, and 3rd-order moments estimation, as well as entropy estimation. We exclude the 1st-order moment, as it can be precisely computed by adding up all updating weights. We conduct two experiments to show the performance under fixed and varying $D:I$ ratios, respectively. We set the relative threshold ε defined in Eq. (2) to 2^{-14}, and the size of the prefilter in each layer to $\frac{\zeta}{12\varepsilon}$ as specified in Section 6.4.

1) We fix the $D:I$ ratio (i.e., $1-\frac{1}{\zeta}$) to 0.5 and evaluate the performance on the CAIDA traces, IMDB, and Yelp reviews datasets. We set the memory allocation to range from 768KB to 2560KB for the first two datasets, and from 5MB to 50MB for the last dataset.

2) We evaluate using the Zipf distribution dataset under 2560KB of memory, with $D:I$ ratio ranging from 0.1 to 0.9, to reflect the scenario where deletions exist and the $D:I$ ratio is bounded.

Setting of the Number of Hierarchical Layers $\ell+1$. Let N be the number of distinct elements in a data stream. Suppose the number of layers in the RUS is $\ell+1$, and the prefilter size on each layer is set to k. On average, the number of elements sampled on the ℓth layer by hierarchical sampling is $N/2^{\ell}$. Then, $N/2^{\ell} \le k$, which means the prefilter on the ℓth layer should be capable of holding all the elements sampled to this layer. For setting the smallest ℓ satisfying this, we have $\ell = \lceil \log_2(N/k) \rceil$. The reason is detailed below.

1) When the number of layers is greater than $\lceil \log_2(N/k) \rceil + 1$, the elements sampled to higher layers can be found in the prefilter on the $\lceil \log_2(N/k) \rceil$th layer. Therefore, these higher layers do not contribute to the moment estimation, as an element e contributes only at its lowest sampled layer $j_b(e)$ by Eqs. (10) and (11). Thus, these layers result in wasted memory, which degrades the HH detection accuracy and reduces the moment estimation accuracy.

2) When the number of layers is fewer than $\lceil \log_2(N/k) \rceil + 1$, according to Eq. (11), the error in moment estimation will be much higher than the real value, and even higher than when the number of layers exceeds $\lceil \log_2(N/k) \rceil + 1$. This is because the information of the HHs, which is expected to be stored in the absent layers required to achieve $\lceil \log_2(N/k) \rceil + 1$ layers, is completely lost.

Due to the deletions and estimation errors, the optimal value in practice may not exactly match $\lceil \log_2(N/k) \rceil + 1$. Therefore, we select the optimal value within $[\lceil \log_2(N/k) \rceil - 1, \lceil \log_2(N/k) \rceil + 3]$ (i.e., $\lceil \log_2(N/k) \rceil + 1 \pm 2$). Due to page limit, we include the results for different number of layers in our technical report [1]. Based on these results, we set the number of layers to 9, 10, 15, and 12 for the CAIDA traces, IMDB, Yelp reviews, and Zipf datasets, respectively.

Performance under Fixed D:I Ratio (Figs. 10a–10l). Compared with UnivMon, the RE of RUS for the 0th-, 2nd-, 3rd-order moment and entropy estimation is 17.2, 2.6×10^3, 1.4×10^6, and 2.1 times lower, respectively, on average under 2560KB of memory for the CAIDA traces and IMDB datasets; it is 1.8, 2.3×10^2, 1.5×10^4, and 5.0 times lower under 50MB of memory for the Yelp reviews dataset.

Analysis of Moment Estimation under Fixed D:I Ratio. (a) The moment estimation error of RUS and Off-RUS shows negligible difference. This is because the OME used by RUS to incrementally update the moment estimation online does not introduce additional error, as described in Section 8.1. (b) The moment estimation error with our RUS is lower than that with UnivMon. This is because in moment estimation, our RUS uses the RAS method to identify HHs and estimate their frequencies, which shows superior performance over the CS+MH method used by UnivMon, as shown in Fig. 8.

Analysis of Moment Estimation under Varying D:I Ratios. (a) As shown in Fig. 11, the estimation error increases with the $D:I$ ratio. This is because Section 9.3.2 indicates that the HH detection accuracy decreases as the $D:I$ ratio increases under a fixed memory allocation, resulting in decreased moment estimation accuracy as indicated by Eq. (11), which calculates the moment by aggregating the weighted frequencies of all HHs. (b) The estimation error of RUS and Off-RUS on the Zipf distribution dataset in the shuffled pattern is lower than that in the sorted pattern. For example, Fig. 11c shows that the estimation error of the 3rd-order moment for RUS on the Zipf distribution dataset in the shuffled pattern can reach as low as 10^{-10}, compared to 10^{-5} in the sorted pattern shown in Fig. 11g. This is because Section 9.3.2 shows that, the HH detection accuracy for the synthetic dataset in the sorted pattern is lower than that in the shuffled pattern, which increases the moment estimation error.

9.5 Processing Speed

We show the average throughput across different memory allocations of methods in per-element frequency estimation (Section 9.2), HH detection (Section 9.3), as well as moment update and estimation (Section 9.4) on three real-world datasets in Figs. 12a–12d.

Analysis of Throughput for Frequency Estimation (Fig. 12a). (a) The throughput of CMS, CS, DS, SuCS, and rSkt is higher than that of RA-SuCS because they use only the integer counters to record frequencies. (b) Compared to RA-CS on three real-world datasets, the throughput of RA-SuCS is 1.7 times higher, because it updates fewer RACs per insertion and deletion compared to RA-CS.

Analysis of Throughput for HH Detection (Fig. 12b). (a) The prefilter-based methods have the lowest throughput compared to the postfilter- and counter-based methods. Specifically, compared with RAS, the throughput of MH+CMS/CS/DS/rSkt is 8.0, 8.2, 8.4, 8.3 times lower. The throughput of CMS/CS/DS/rSkt+MH is 2.9, 2.9, 3.2, 2.9 times lower, and the throughput of SS$^{\pm}$ is 2.4 times lower. This is because the prefilter-based methods must query for the element from the prefilter once an element carried by a stream tuple arrives, with $O(k)$ lookup time cost for the min-heap. For the postfilter-based methods, the element is queried from the postfilter only when it is regarded as a HH, with $O(k)$ lookup time cost for the min-heap. For the counter-based method, SS$^{\pm}$, it achieves higher throughput as it maintains a hash table to accelerate the query speed of the min-heap. However, the min-heap is relatively large due to the *single element-frequency table design*, so the sift-up and sift-down operations, which cost $O(\log k)$, are still significant. (b) RAS shows the highest throughput, since we use KP-CF as a prefilter, which offers $O(1)$ query and update time costs, making the overall throughput of RAS higher than that of the comparison methods.

Analysis of Throughput for Moment Estimation (Figs. 12c and 12d). (a) The update throughput of RUS is higher (i.e., 3.2 times higher) than that of UnivMon. This is because for each arrival tuple, both algorithms have to update multiple HH detectors, and the HH detector used by RUS, namely RAS, offers higher update throughput than that of UnivMon, i.e., CS+MH. (b) The update throughput of RUS is slightly lower (i.e., 1.1 times lower) than that of Off-RUS. This

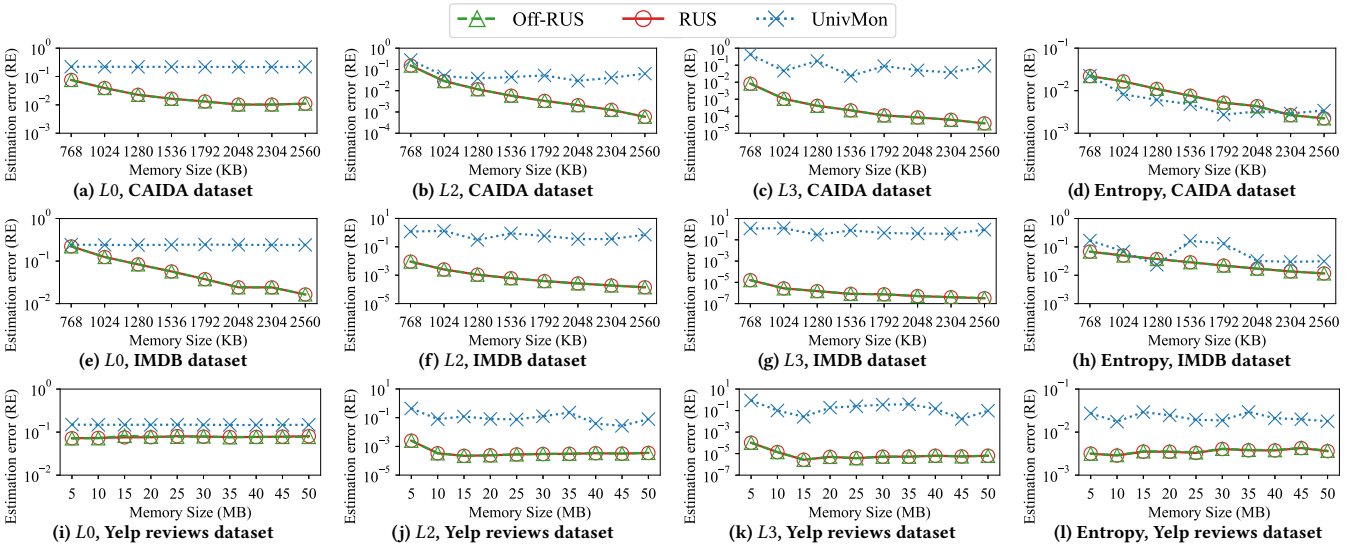


Figure 10: Moment estimation errors when D:I ratio is 0.5

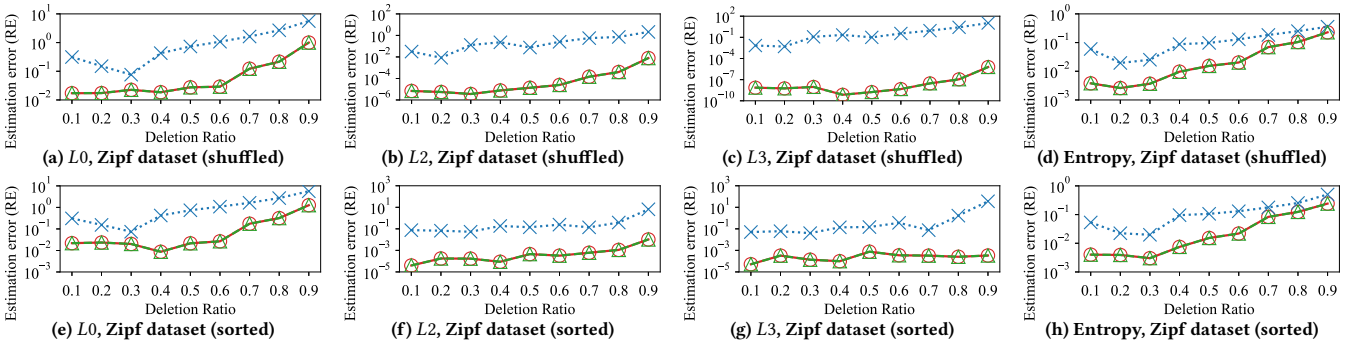


Figure 11: Moment estimation errors under varying $D:I$ ratios

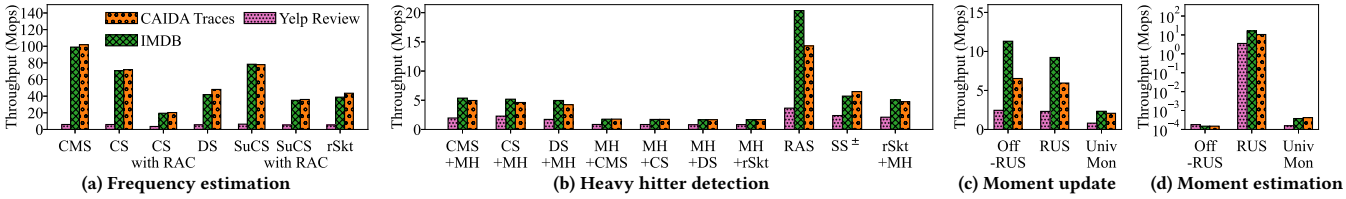


Figure 12: Throughput of methods involving per-element frequency estimation, HH detection, and moment estimation

is because RUS utilizes OME, which needs to update the moment estimation during the handling of the data stream tuples. (c) For the same reason, RUS achieves significantly higher throughput for moment estimation than that of Off-RUS and UnivMon. Specifically, the throughput of RUS for moment estimation is 3.0×10^4 and 6.6×10^4 times higher than that of UnivMon and Off-RUS, respectively. Note that Off-RUS and UnivMon utilize the *recursive summation technique*, which requires a linear scan of all the HHs from the highest layer downwards to the lowest layer upon queries.

10 CONCLUSION AND FUTURE WORK

For a data stream in the *bounded deletion model*, we propose RUS to simultaneously collect multiple kinds of statistics online: per-element frequency, HHs, frequency moments, and frequency distribution. Three main components of our RUS are the RAS to detect HHs in the *bounded deletion stream*, the OME to estimate the frequency moments online, and the 16-bit RAC to further improve the frequency estimation accuracy. Our experiments show that our RUS improves the F1 Score of detecting HHs by 16% ~ 69%, and increases the throughput of frequency moment estimation by 3.0×10^4 times.

In the future, we plan to adapt our solution to various scenarios described in Section 5.2, including network measurement, pattern mining in text corpora, and database query optimization. For network measurement, we will focus on exploring ways to apply multiple RUSs to collect statistics in distributed settings [80]. For pattern mining, we will extend RUS from mining frequent n-grams to the more general case of frequent sequential patterns [81]. For database query optimization, we will focus on integrating our RUS with the data-driven cardinality estimation model [38, 71] using statistics across multiple relations, to determine the cardinality of queries.

REFERENCES

[1] [n. d.]. A Universal Sketch for Estimating Heavy Hitters and Per-Element Frequency Moments in Data Streams with Bounded Deletions [technical report]. https://anonymous.4open.science/r/removable-universal-sketch/technical_report/TechnicalReport.pdf.

[2] Sugam Agarwal, Murali Kodialam, and TV Lakshman. 2013. Traffic engineering in software defined networks. In *Proc. IEEE INFOCOM*. 2211–2219.

[3] Noga Alon, Phillip B Gibbons, Yossi Matias, and Mario Szegedy. 1999. Tracking join and self-join sizes in limited storage. In *Proc. of ACM PODS*. 10–20.

[4] Foteini Alvanaki and Sebastian Michel. 2014. Tracking set correlations at large scale. In *Proc. of ACM SIGMOD*. 1507–1518.

[5] Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. 2012. A method of moments for mixture models and hidden Markov models. In *Proc. of the ACM COLT*, Vol. 23. 33.1–33.34.

[6] Vladimir Braverman, Stephen R Chestnut, Nikita Ivkin, Jelani Nelson, Zhengyu Wang, and David P Woodruff. 2017. Bptree: An l2 heavy hitters algorithm using constant memory. In *Proc. of ACM PODS*. 361–376.

[7] Vladimir Braverman and Rafail Ostrovsky. 2013. Generalizing the layering method of indyk and woodruff: Recursive sketches for frequency-based vectors on streams. In *Proc. of APPROX Workshop*, Vol. 8096. 58–70.

[8] Vladimir Braverman, David Woodruff, and Lin Yang. 2018. Revisiting Frequency Moment Estimation in Random Order Streams. In *Proc. of ICALP*, Vol. 107. 25:1–25:14.

[9] CAIDA. 2016. The CAIDA Anonymized Internet Traces. http://www.caida.org/data/overview/

[10] Moses Charikar, Kevin Chen, and Martin Farach-Colton. 2004. Finding frequent items in data streams. *Theoretical Computer Science* 312, 1 (2004), 3–15.

[11] Peiqing Chen, Dong Chen, Lingxiao Zheng, Jizhou Li, and Tong Yang. 2021. Out of many we are one: Measuring item batch with clock-sketch. In *Proc. of ACM SIGMOD*. 261–273.

[12] Zhida Chen, Gao Cong, and Walid G Aref. 2020. STAR: A distributed stream warehouse system for spatial data. In *Proc. of ACM SIGMOD*. 2761–2764.

[13] Michael P Connolly, Nicholas J Higham, and Theo Mary. 2021. Stochastic rounding and its probabilistic backward error analysis. *SIAM Journal on Scientific Computing* 43, 1 (2021), A566–A585.

[14] Jeffrey Considine, Feifei Li, George Kollios, and John Byers. 2004. Approximate aggregation techniques for sensor databases. In *Proc. of IEEE ICDE*. 449–460.

[15] Graham Cormode. 2022. Current trends in data summaries. *ACM SIGMOD Record* 50, 4 (2022), 6–15.

[16] Graham Cormode and Shan Muthukrishnan. 2005. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms* 55, 1 (2005), 58–75.

[17] Hetong Dai, Heng Li, Che-Shao Chen, Weiyi Shang, and Tse-Hsun Chen. 2020. Logram: Efficient log parsing using n n-gram dictionaries. *IEEE Transactions on Software Engineering* 48, 3 (2020), 879–892.

[18] Kyle B Deeds, Dan Suciu, and Magdalena Balazinska. 2023. SafeBound: A Practical System for Generating Cardinality Bounds. In *Proc. of ACM SIGMOD*, Vol. 1. 1–26.

[19] Cristian Estan and George Varghese. 2002. New directions in traffic measurement and accounting. In *Proc. of ACM SIGCOMM*. 323–336.

[20] Cristian Estan, George Varghese, and Mike Fisk. 2003. Bitmap algorithms for counting active flows on high speed links. In *Proc. of ACM SIGCOMM*. 153–166.

[21] Bin Fan, Dave G Andersen, Michael Kaminsky, and Michael D Mitzenmacher. 2014. Cuckoo filter: Practically better than bloom. In *Proc. of ACM CoNEXT*. 75–88.

[22] Zhuochen Fan, Ruixin Wang, Yalun Cai, Ruwen Zhang, Tong Yang, Yuhan Wu, Bin Cui, and Steve Uhlig. 2023. OneSketch: A Generic and Accurate Sketch for Data Streams. *IEEE Transactions on Knowledge and Data Engineering* 35, 12 (2023), 12887–12901.

[23] Edward Gan, Jialin Ding, Kai Sheng Tai, Vatsal Sharan, and Peter Bailis. 2018. Moment-Based Quantile Sketches for Efficient High Cardinality Aggregation Queries. In *Proc. of VLDB Endow.*, Vol. 11. 1647–1660.

[24] Minos Garofalakis, Johannes Gehrke, and Rajeev Rastogi. 2016. *Data stream management: processing high-speed data streams*. Springer.

[25] Michael Geller and Pramod Nair. 2018. 5G security innovation with Cisco. *Whitepaper Cisco Public* (2018), 1–29.

[26] Stefan Heule, Marc Nunkesser, and Alexander Hall. 2013. Hyperloglog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm. In *Proc. of Springer EDBT*. 683–692.

[27] He Huang, Jiakun Yu, Yang Du, Jia Liu, Haipeng Dai, and Yu-E Sun. 2023. Memory-Efficient and Flexible Detection of Heavy Hitters in High-Speed Networks. In *Proc. of ACM SIGMOD*, Vol. 1. 1–24.

[28] Yesdaulet Izenov, Asoke Datta, Florin Rusu, and Jun Hyung Shin. 2021. COMPASS: Online sketch-based query optimization for in-memory databases. In *Proc. of ACM SIGMOD*. 804–816.

[29] Rajesh Jayaram and David P Woodruff. 2018. Data streams with bounded deletions. In *Proc. of ACM PODS*. 341–354.

[30] Piotr Jurkiewicz, Grzegorz Rzym, and Piotr Boryło. 2021. Flow length and size distributions in campus Internet traffic. *Computer Communications* 167 (2021), 15–30.

[31] Aarati Kakaraparthy, Jignesh M Patel, Brian P Kroth, and Kwanghyun Park. 2022. VIP hashing: adapting to skew in popularity of data on the fly. In *Proc. of VLDB Endow.*, Vol. 15. 1978–1990.

[32] Abhishek Kumar, Minho Sung, Jun Xu, and Jia Wang. 2004. Data streaming algorithms for efficient and accurate estimation of flow size distribution. *ACM SIGMETRICS Performance Evaluation Review* 32, 1 (2004), 177–188.

[33] Ashwin Lall, Vyas Sekar, Mitsunori Ogihara, Jun Xu, and Hui Zhang. 2006. Data streaming algorithms for estimating entropy of network traffic. *ACM SIGMETRICS Performance Evaluation Review* 34, 1 (2006), 145–156.

[34] Kasper Green Larsen, Rasmus Pagh, and Jakub Tětek. 2021. Countsketches, feature hashing and the median of three. In *International Conference on Machine Learning*. PMLR, 6011–6020.

[35] Alexandru Lavric and Valentin Popa. 2017. Internet of things and LoRa™ low-power wide-area networks: a survey. In *Proc. of IEEE ISSCS*. 1–5.

[36] Viktor Leis, Andrey Gubichev, Atanas Mirchev, Peter Boncz, Alfons Kemper, and Thomas Neumann. 2015. How good are query optimizers, really?. In *Proc. of VLDB Endow.*, Vol. 9. 204–215.

[37] Jizhou Li, Zikun Li, Yifei Xu, Shiqi Jiang, Tong Yang, Bin Cui, Yafei Dai, and Gong Zhang. 2020. Wavingsketch: An unbiased and generic sketch for finding top-k items in data streams. In *Proc. of ACM SIGKDD*. 1574–1584.

[38] Pengfei Li, Wenqing Wei, Rong Zhu, Bolin Ding, Jingren Zhou, and Hua Lu. 2023. ALECE: An Attention-based Learned Cardinality Estimator for SPJ Queries on Dynamic Workloads. In *Proc. of VLDB Endow.*, Vol. 17. 197–210.

[39] Weihe Li and Paul Patras. 2024. Stable-Sketch: A Versatile Sketch for Accurate, Fast, Web-Scale Data Stream Processing. In *Proc. of ACM WWW*. 4227–4238.

[40] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *Proc. of ACM SIGMOD*. 1729–1744.

[41] Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. 2022. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *Proc. of IEEE INFOCOM*. IEEE, 1749–1758.

[42] Zaoxing Liu, Antonis Manousis, Gregory Vorsanger, Vyas Sekar, and Vladimir Braverman. 2016. One sketch to rule them all: Rethinking network flow monitoring with univmon. In *Proc. of ACM SIGCOMM*. 101–114.

[43] Qiang Long, Wei Wang, Jinfu Deng, Song Liu, Wenhao Huang, Fangying Chen, and Sifan Liu. 2019. A distributed system for large-scale n-gram language models at Tencent. In *Proc. of VLDB Endow.*, Vol. 12. 2206–2217.

[44] Jiaheng Lu, Chunbin Lin, Wei Wang, Chen Li, and Haiyong Wang. 2013. String similarity measures and joins with synonyms. In *Proc. of ACM SIGMOD*. 373–384.

[45] Nishad Manerikar and Themis Palpanas. 2009. Frequent items in streaming data: An experimental evaluation of the state-of-the-art. *Data & Knowledge Engineering* 68, 4 (2009), 415–430.

[46] Gurmeet Singh Manku and Rajeev Motwani. 2002. Approximate frequency counts over data streams. In *Proc. of VLDB Endow.* 346–357.

[47] Dimitrios Melissourgos, Haibo Wang, Shigang Chen, Chaoyi Ma, and Shiping Chen. 2023. Single Update Sketch with Variable Counter Structure. In *Proc. of VLDB Endow.*, Vol. 16. 4296–4309.

[48] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. 2005. Efficient computation of frequent and top-k elements in data streams. In *Proc. of Springer ICDT*, Vol. 3363. 398–412.

[49] Jayanta Mondal and Amol Deshpande. 2014. Eagr: Supporting continuous ego-centric aggregate queries over large dynamic graphs. In *Proc. of ACM SIGMOD*. 1335–1346.

[50] Shanmugavelayutham Muthukrishnan et al. 2005. Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science* 1, 2 (2005), 117–236.

[51] Dang Nguyen, Wei Luo, Tu Dinh Nguyen, Svetha Venkatesh, and Dinh Phung. 2019. Sqn2vec: Learning sequence representation via sequential patterns with a gap constraint. In *Proc. of ECML PKDD*. 569–584.

[52] Thanh Tam Nguyen, Thanh Trung Huynh, Hongzhi Yin, Matthias Weidlich, Thanh Thi Nguyen, Thai Son Mai, and Quoc Viet Hung Nguyen. 2023. Detecting rumours with latency guarantees using massive streaming data. *The VLDB Journal* 32, 2 (2023), 369–387.

[53] Stuart L Pardau. 2018. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y* 23 (2018), 68.

[54] Debjyoti Paul, Yanqing Peng, and Feifei Li. 2019. Bursty event detection throughout histories. In *Proc. of IEEE ICDE*. 1370–1381.

[55] Pratanu Roy, Arijit Khan, and Gustavo Alonso. 2016. Augmented sketch: Faster and more accurate stream processing. In *Proc. of ACM SIGMOD*. 1449–1463.

[56] Tony Saad and Giovanna Ruai. 2019. PyMaxEnt: A Python software for maximum entropy moment reconstruction. *SoftwareX* 10 (2019), 100353.

[57] Khaled Mohammed Saifuddin, Corey May, Farhan Tanvir, Muhammad Ifte Khairul Islam, and Esra Akbas. 2023. Seq-HyGAN: Sequence Classification via Hypergraph Attention Network. In *Proc. of ACM CIKM*. 2167–2177.

[58] P Griffiths Selinger, Morton M Astrahan, Donald D Chamberlin, Raymond A Lorie, and Thomas G Price. 1979. Access path selection in a relational database

management system. In *Proc. of ACM SIGMOD*. 23–34.
[59] HyungBin Seo and MyungKeun Yoon. 2023. Generative intrusion detection and prevention on data stream. In *Proc. of USENIX Security*. 4319–4335.
[60] Cha Hwan Song, Pravein Govindan Kannan, Bryan Kian Hsiang Low, and Mun Choon Chan. 2020. FCM-Sketch: Generic Network Measurements with Data Plane Support. In *Proc. of ACM CoNEXT*. 78–92.
[61] Zehua Sun, Huanqi Yang, Kai Liu, Zhimeng Yin, Zhenjiang Li, and Weitao Xu. 2022. Recent advances in LoRa: A comprehensive survey. *ACM Transactions on Sensor Networks* 18, 4 (2022), 1–44.
[62] Kai Sheng Tai, Vatsal Sharan, Peter Bailis, and Gregory Valiant. 2018. Sketching linear classifiers over data streams. In *Proc. of ACM SIGMOD*. 757–772.
[63] Lu Tang, Qun Huang, and Patrick PC Lee. 2019. MV-Sketch: A Fast and Compact Invertible Sketch for Heavy Flow Detection in Network Data Streams. In *Proc. of IEEE INFOCOM*. 2026–2034.
[64] Daniel Ting. 2018. Data Sketches for Disaggregated Subset Sum and Frequent Item Estimation. In *Proc. of ACM SIGMOD*. 1129–1140.
[65] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.
[66] Haibo Wang, Chaoyi Ma, Olufemi O Odegbile, Shigang Chen, and Jih-Kwon Peir. 2021. Randomized error removal for online spread estimation in data streaming. In *Proc. of VLDB Endow.*, Vol. 14. 1040–1052.
[67] Haibo Wang, Chaoyi Ma, Olufemi O Odegbile, Shigang Chen, and Jih-Kwon Peir. 2022. Randomized Error Removal for Online Spread Estimation in High-Speed Networks. *IEEE/ACM TON* 31, 2 (2022), 558–573.
[68] Larry Wasserman. 2004. *All of statistics: a concise course in statistical inference*. Vol. 26. Springer.
[69] Kyu-Young Whang, Brad T Vander-Zanden, and Howard M Taylor. 1990. A linear-time probabilistic counting algorithm for database applications. *ACM Transactions on Database Systems* 15, 2 (1990), 208–229.
[70] David P Woodruff and Samson Zhou. 2021. Separations for Estimating Large Frequency Moments on Data Streams. In *Proc. of ICALP*, Vol. 198. 112:1–112:21.
[71] Ziniu Wu, Parimarjan Negi, Mohammad Alizadeh, Tim Kraska, and Samuel Madden. 2023. FactorJoin: a new cardinality estimation framework for join queries. In *Proc. of ACM SIGMOD*. 1–27.
[72] Qingjun Xiao, Xuyuan Cai, Yifei Qin, Zhiying Tang, Shigang Chen, and Yu Liu. 2023. Universal and Accurate Sketch for Estimating Heavy Hitters and Moments in Data Streams. *IEEE/ACM TON* 31, 5 (2023), 1919–1934.
[73] Qingjun Xiao, Yuexiao Cai, Yunpeng Cao, and Shigang Chen. 2023. Accurate and O(1)-Time Query of Per-Flow Cardinality in High-Speed Networks. *IEEE/ACM TON* 31, 6 (2023), 2994–3009.
[74] Qingjun Xiao, Shigang Chen, You Zhou, Min Chen, Junzhou Luo, Tengli Li, and Yibei Ling. 2017. Cardinality estimation for elephant flows: A compact solution based on virtual register sharing. *IEEE/ACM TON* 25, 6 (2017), 3738–3752.
[75] Qingjun Xiao, Yifei Li, and Yeke Wu. 2023. Finding recently persistent flows in high-speed packet streams based on cuckoo filter. *Computer Networks* 237 (2023), 110097.
[76] Qingjun Xiao, Zhiying Tang, and Shigang Chen. 2020. Universal online sketch for tracking heavy hitters and estimating moments of data streams. In *Proc. of IEEE INFOCOM*. 974–983.
[77] Qingjun Xiao, Haotian Wang, and Guannan Pan. 2022. Accurately Identify Time-decaying Heavy Hitters by Decay-aware Cuckoo Filter along Kicking Path. In *Proc. of IEEE/ACM IWQoS*. 1–10.
[78] Qingjun Xiao, You Zhou, and Shigang Chen. 2017. Better with fewer bits: Improving the performance of cardinality estimation of large data streams. In *Proc. of IEEE INFOCOM*. 1–9.
[79] Guorui Xie, Qing Li, Guanglin Duan, Yong Jiang, Zhuyun Qi, Shuo Liu, and Qiaoling Wang. 2023. Efficient Flow Recording with InheritSketch on Programmable Switches. In *2023 IEEE 43rd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1–11.
[80] Yuchen Xu, Wenfei Wu, Bohan Zhao, Tong Yang, and Yikai Zhao. 2023. MimoSketch: A Framework to Mine Item Frequency on Multiple Nodes with Sketches. In *Proc. of ACM SIGKDD*. 2838–2849.
[81] Da Yan, Wenwen Qu, Guimu Guo, Xiaoling Wang, and Yang Zhou. 2022. PrefixFPM: a parallel framework for general-purpose mining of frequent and closed patterns. *The VLDB Journal* 31, 2 (2022), 253–286.
[82] Kaicheng Yang, Sheng Long, Qilong Shi, Yuanpeng Li, Zirui Liu, Yuhan Wu, Tong Yang, and Zhengyi Jia. 2023. SketchINT: Empowering int with towersketch for per-flow per-switch measurement. *IEEE Transactions on Parallel and Distributed Systems* 34, 11 (2023), 2876–2894.
[83] Mingran Yang, Junbo Zhang, Akshay Gadre, Zaoxing Liu, Swarun Kumar, and Vyas Sekar. 2020. Joltik: enabling energy-efficient" future-proof" analytics on low-power wide-area networks. In *Proc. of ACM MobiCom*. 1–14.
[84] Tong Yang, Siang Gao, Zhouyi Sun, Yufei Wang, Yulong Shen, and Xiaoming Li. 2019. Diamond sketch: Accurate per-flow measurement for big streaming data. *IEEE Transactions on Parallel and Distributed Systems* 30, 12 (2019), 2650–2662.
[85] Tong Yang, Junzhi Gong, Haowei Zhang, Lei Zou, Lei Shi, and Xiaoming Li. 2018. Heavyguardian: Separate and guard hot items in data streams. In *Proc. of ACM SIGKDD*. 2584–2593.
[86] Tong Yang, Jie Jiang, Peng Liu, Qun Huang, Junzhi Gong, Yang Zhou, Rui Miao, Xiaoming Li, and Steve Uhlig. 2018. Elastic sketch: Adaptive and fast network-wide measurements. In *Proc. of ACM SIGCOMM*. 561–575.
[87] Tong Yang, Haowei Zhang, Jinyang Li, Junzhi Gong, Steve Uhlig, Shigang Chen, and Xiaoming Li. 2019. HeavyKeeper: an accurate algorithm for finding Top-k elephant flows. *IEEE/ACM TON* 27, 5 (2019), 1845–1858.
[88] Yelp. 2021. Yelp Open Dataset. https://www.yelp.com/dataset/
[89] Quanwei Zhang, Qingjun Xiao, and Yuexiao Cai. 2023. A generic sketch for estimating super-spreaders and per-flow cardinality distribution in high-speed data streams. *Computer Networks* 237 (2023), 110059.
[90] Fuheng Zhao, Divyakant Agrawal, Amr El Abbadi, and Ahmed Metwally. 2022. SpaceSaving$^{\pm}$: an optimal algorithm for frequency estimation and frequent items in the bounded-deletion model. In *Proc. of VLDB Endow.*, Vol. 15. 1215–1227.
[91] Fuheng Zhao, Punnal Ismail Khan, Divyakant Agrawal, Amr El Abbadi, Arpit Gupta, and Zaoxing Liu. 2023. Panakos: Chasing the Tails for Multidimensional Data Streams. In *Proc. of VLDB Endow.*, Vol. 16. 1291–1304.
[92] Fuheng Zhao, Sujaya Maiyya, Ryan Wiener, Divyakant Agrawal, and Amr El Abbadi. 2021. KLL$^{\pm}$approximate quantile sketches over dynamic datasets. In *Proc. of VLDB Endow.*, Vol. 14. 1215–1227.
[93] Yang Zhou, Tong Yang, Jie Jiang, Bin Cui, Minlan Yu, Xiaoming Li, and Steve Uhlig. 2018. Cold filter: A meta-framework for faster and more accurate stream processing. In *Proc. of ACM SIGMOD*. 741–756.
[94] You Zhou, Yian Zhou, Shigang Chen, and Youlin Zhang. 2018. Highly compact virtual active counters for per-flow traffic measurement. In *Proc. of IEEE INFOCOM*. 1–9.

A COMPARISON OF TIME COMPLEXITY

The time cost of the methods supporting deletions is given in Table 3.

Table 3: Time complexity of algorithms support deletions

Task	Method	Update	Query
HH Detection	SS$^{\pm}$	$O(\log(filter\ size))$	$O(1)$
	Sketch + Min-heap	$O(1)$	$O(filter\ size)$
	Min-heap + Sketch	$O(filter\ size)$	$O(filter\ size)$
	Our RAS	$O(1)$	$O(1)$
Moment Estimation	UnivMon	$O(number\ of\ layers)$	$O(filter\ size)$
	Our RUS	$O(1)$	$O(1)$

B UPDATE PROCEDURE OF RAC

B.1 Pseudocode For Updating RAC

Suppose an RAC C needs to be added by a weight w, i.e., $C = C + w$, where w is allowed to be either a positive or a negative integer. We present the pseudocode of updating the RAC C in Algorithm 2.

Rounding of Increment. Since the counter value $V(C)$ has been multiplied by $2^{C.\alpha}$ in Eq. (7), we need to divide the increment w by $2^{C.\alpha}$, so that it can be added to the coefficient $C.\beta$. However, the division result $w\ /\ 2^{C.\alpha}$ may not be an integer. So at Line 1, we call the $\texttt{roundedDiv}(w, 2^{C.\alpha})$ to calculate an integer division result.

The function $\texttt{roundedDiv}(w, v)$ is defined at Line 10. It returns an integer result randomly alternating between $\lfloor w\ /\ v \rfloor$ and $\lceil w\ /\ v \rceil$, which ensures the expected value is equal to $w\ /\ v$. Inside the function $\texttt{roundedDiv}(w, v)$, we firstly check whether w is a negative value, and if yes, we return $-\texttt{roundedDiv}(-w, v)$ at Line 11. Then, assuming w is a positive value, we divide w by v, and obtain the quotient $\lfloor w\ /\ v \rfloor$ and the remainder $(w \bmod v)$ at Line 12. To reduce the rounding error of $(w \bmod v)\ /\ v$, we use a *stochastic rounding technique* [13]: At Line 13, we generate a random integer $r \in [0,\ v)$. If r is smaller than $(w \bmod v)$, the result is added by 1.

Initial Update. At Line 1, we define a local variable $\tilde{\beta}$ to store the updated value of the coefficient. Here, we add together the

Algorithm 2: Update Procedure of RAC

Input: RAC C, increment w as a signed integer

1 $\tilde{\beta} = 2^{\mathcal{L}_\beta} + C.\beta + (2\,C.\rho - 1) \cdot \texttt{roundedDiv}(w,\, 2^{C.\alpha}), \quad \tilde{\alpha} = C.\alpha$
2 **if** $\tilde{\beta} == 0$ **then** $C.\rho = 1 - C.\rho, \quad \tilde{\beta} = 2^{\mathcal{L}_\beta}, \quad \tilde{\alpha} = 1$
3 **if** $\tilde{\beta} \cdot 2^{\tilde{\alpha}} - 2^{\mathcal{L}_\beta} < 0$ **then**
4 $\quad C.\rho = 1 - C.\rho, \quad \tilde{\beta} = \texttt{roundedDiv}(2^{\mathcal{L}_\beta + 1} - \tilde{\beta} \cdot 2^{\tilde{\alpha}},\, 2^{\tilde{\alpha}})$
5 **while** $\tilde{\beta} < 2^{\mathcal{L}_\beta} \ \vee \ \tilde{\beta} \geq 2^{\mathcal{L}_\beta + 1}$ **do**
6 $\quad \Delta\alpha = \mathcal{L}_\beta - \lfloor \log_2(\tilde{\beta}) \rfloor, \quad \tilde{\alpha} = \tilde{\alpha} - \Delta\alpha$
7 $\quad$ **if** $\Delta\alpha > 0$ **then** $\tilde{\beta} = \tilde{\beta} \cdot 2^{\Delta\alpha}$ **else** $\tilde{\beta} = \texttt{roundedDiv}(\tilde{\beta},\, 2^{-\Delta\alpha})$
8 **if** $\tilde{\alpha} \geq 2^{\mathcal{L}_\alpha}$ **then throw** *OverflowException*
9 **else** $C.\beta = \tilde{\beta} - 2^{\mathcal{L}_\beta}, \quad C.\alpha = \tilde{\alpha}$
10 **Function** roundedDiv(w, v): // dividend, divisor
11 $\quad$ **if** $w < 0$ **then return** $-\texttt{roundedDiv}(-w, v)$
12 $\quad$ **if** $w \geq v$ **then return** $\lfloor w/v \rfloor + \texttt{roundedDiv}(w \bmod v, v)$
13 $\quad$ **else return** $\mathbf{1}_{r<w}$, where r is a random integer in $[0, v)$

implicit leading 1 bit $2^{\mathcal{L}_\beta}$, the stored coefficient value $C.\beta$, and the rounded increment $\texttt{roundedDiv}(w, 2^{C.\alpha})$, multiplied by the sign of the counter $(2\,C.\rho - 1)$. At Line 1, we also define a local variable $\tilde{\alpha}$ to hold the exponent $C.\alpha$. At Line 2, we handle an exceptional case that $\tilde{\beta}$ becomes 0, since Lines 5–7 assume $\tilde{\beta}$ is not equal to 0. If it happens, Eq. (7) implies that $V(C)$ equals $-(2\,C.\rho - 1) \cdot 2^{\mathcal{L}_\beta}$. Thus, we flip the sign bit $C.\rho$, update the coefficient $\tilde{\beta}$ to $2^{\mathcal{L}_\beta}$, and update the exponent $\tilde{\alpha}$ to 1. Next, at Line 3, we check whether $\tilde{\beta} \cdot 2^{\tilde{\alpha}} - 2^{\mathcal{L}_\beta}$ is negative, implying $V(C)/(2\,C.\rho - 1)$ is negative in Eq. (7). To ensure this term is non-negative, at Line 4, we flip the sign bit $C.\rho$ and update $\tilde{\beta}$ to $\texttt{roundedDiv}(2^{\mathcal{L}_\beta+1} - \tilde{\beta} \cdot 2^{\tilde{\alpha}}, 2^{\tilde{\alpha}})$, whose expected value is $\frac{2^{\mathcal{L}_\beta+1} - \tilde{\beta} \cdot 2^{\tilde{\alpha}}}{2^{\tilde{\alpha}}}$. Since $\frac{2^{\mathcal{L}_\beta+1} - \tilde{\beta} \cdot 2^{\tilde{\alpha}}}{2^{\tilde{\alpha}}} \cdot 2^{\tilde{\alpha}} - 2^{\mathcal{L}_\beta} = -\left(\tilde{\beta} \cdot 2^{\tilde{\alpha}} - 2^{\mathcal{L}_\beta}\right)$, the counter value $V(C)$ is unchanged by Eq. (7).

Correct Over/Underflow. At Line 5, we check $\tilde{\beta} < 2^{\mathcal{L}_\beta}$ or $\tilde{\beta} \geq 2^{\mathcal{L}_\beta+1}$, to detect the underflow or overflow of the coefficient $\tilde{\beta}$, which has only $\mathcal{L}_\beta$ bits and an implicit leading 1 bit. If it happens, at Line 6, we calculate the exponent increment $\Delta\alpha$ by $\mathcal{L}_\beta - \lfloor \log_2(\tilde{\beta}) \rfloor$, and update the local variable for exponent $\tilde{\alpha}$ accordingly. Then, at Line 7, if $\Delta\alpha > 0$, we can multiply $\tilde{\beta}$ by $2^{\Delta\alpha}$ to correct its underflow. Otherwise, since $\Delta\alpha < 0$, we divide $\tilde{\beta}$ by $2^{-\Delta\alpha}$ to correct its overflow. Here, we have to reuse the *stochastic rounding technique*. However, $\texttt{roundedDiv}(\tilde{\beta},\, 2^{-\Delta\alpha})$ may be equal to $2^{\mathcal{L}_\beta+1}$, causing $\tilde{\beta}$ overflows again. So we check it by the **while** statement at Line 5, and correct it at Lines 6–7. Finally, Line 8 checks if the exponent $\tilde{\alpha}$ overflows. If not, at Line 9, we assign $\tilde{\beta} - 2^{\mathcal{L}_\beta}$ to $C.\beta$, and assign $\tilde{\alpha}$ to $C.\alpha$.

B.2 Another Example of Updating RAC

In Fig. 6, we show a general case to update RAC. In this subsection, we show an exceptional case that requires special handling. Assume an RAC C with $\rho = 1b$ (denoted by +), $\alpha = 001b$, $\tilde{\beta} = ①001b$, and $\mathcal{L}_\beta = 3$, where ① represents the implicit leftmost 1 bit. Then, we assume w is $-18 = -1 \cdot 2^{001b} \cdot 1001b$ in this example. Since the counter value $V(C)$ has been multiplied by $2^{C.\alpha}$ in Eq. (7), we need to divide the increment w by $2^{C.\alpha}$, so that it can be added to the coefficient $C.\beta$. We denote the division result as $\Delta\tilde{\beta}$, yielding $-1 \cdot 1001b$. We show the case when $\Delta\tilde{\beta}$ is $-1 \cdot 1001b$ by two phases.

- **Obtain new $\tilde{\beta}$.** As shown in Fig. 13, we obtain the new value of $\tilde{\beta}$ by adding $\Delta\tilde{\beta}$ to $\tilde{\beta}$, resulting in $-1 \cdot 1001b + ①001b = ⓪000b$.
- **Reinitialize.** Since the value $V(C)$ in Eq. (7) becomes $-2^{\mathcal{L}_\beta}$ when $\tilde{\beta}$ is 0, we proceed to directly reinitialize ρ, α, and $\tilde{\beta}$ to $0b$, $001b$, and $①000b$, respectively, ensuring the representation of RAC C is maintained while avoiding the underflow of $\tilde{\beta}$.

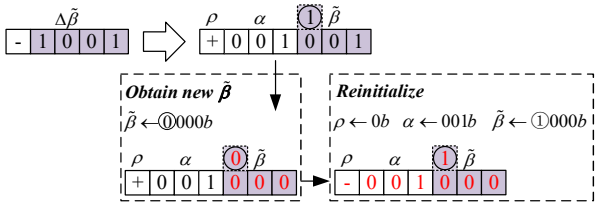


Figure 13: An exceptional case of updating an RAC C

C PSEUDOCODE FOR OME

We present the procedure of our OME in Algorithm 3 based on the functions given in Algorithm 1. As a stream tuple (e, w) arrives, Line 1 reuses the symbol j_t as the topmost sampled layer of the element e for brevity. The rest of the pseudocode consists of two phases: Firstly, from the j_tth layer downwards to the 0th layer, we propagate the element e downwards and track its *bottom layer* to incrementally update the moment estimation by Eq. (11). Secondly, during the propagation, for the elements named e' that are no longer HHs and have to be swapped out, we remove their contribution to moment calculation by Eq. (9) to update the moment estimation.

Update Moment by e. We define two variables, j_b^{old} and j_b^{new} at Line 1, to track the old and new *bottom layers* of the element e, respectively. Then, we perform HH propagation for each jth layer with $j \in [j_t, \ldots, 0]$, as shown in Line 2. Lines 3–13 is almost the same as Lines 2–8 in Algorithm 1, except that we keep e's current frequency $\hat{f}_e^{new}$ at Lines 5, 8 and 11 when j equals j_t, in order to propagate the element e and its current frequency $\hat{f}_e^{new}$ to lower layers. At Line 6, we also update j_b^{old} to the current layer j if e exists in the prefilter on the jth layer, to track the old *bottom layer* of e.

At Line 14, if the kicked element e' returned at Line 13 is e (i.e., the condition at Line 24 of Algorithm 1 holds), the insertion fails. So we terminate the propagation procedure by command **break**. Otherwise, if e' is not e, we proceed with the propagation at Lines 15–18. Then, if we reach Line 19, the element e is successfully inserted into the prefilter on the current layer j, so we update j_b^{new} to j. After that, outside the loop, we have j_b^{old} pointing to the old *bottom layer* where e is already contained by the prefilter there before the tuple (e, w) arrival, while j_b^{new} is the new *bottom layer* pointing to the last layer where e is successfully inserted into a prefilter.

At Line 20, using OME expressed in Eq. (11), we incrementally update the estimation of the moment $\hat{L}_0$ by adding the new contribution and removing the old contribution of the element e. The former is represented by $\mathbf{1}_{j_b^{new} \leq j_t} \cdot 2^{j_b^{new}} \cdot g(\hat{f}_e^{new})$, where $\mathbf{1}_{j_b^{new} \leq j_t}$ indicates if the element e exists in at least one of the prefilters after

Algorithm 3: OME's Online Estimation Procedure

Input: an arrival tuple consisting of element ID e and weight w
State: prefilter R_j and sketch M_j on each layer $j \in [0, \ell]$

1 $j_t = \min\big(\ell, \arg\max_j \{\wedge_{1\le i\le j} h_i(e) \neq 0\}\big)$, $j_b^{new} = j_b^{old} = j_t + 1$
2 **foreach** $j = j_t, j_t - 1, \ldots, 0$ **do**
3 $\langle u, v\rangle$ = `lookupFilter`(R_j, e), $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle$=$\langle$**NIL**, 0, 0$\rangle$
4 **if** $u \neq$ **NIL** $\wedge\ v \neq$ **NIL** $\wedge\ R_j.B[u][v].id == e$ **then**
5 **if** $j == j_t$ **then** $\hat{f}_e^{new} = w + R_{j_t}.B[u][v].newFreq$
6 `updateFilter`$(R_j, u, v, \hat{f}_e^{new})$, $j_b^{old} = j$
7 **else if** $u \neq$ **NIL** $\wedge\ v \neq$ **NIL** **then**
8 **if** $j == j_t$ **then** $\hat{f}_e^{new} = w$, $\hat{f}_e^{old} = 0$
9 $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle$ = `insertFilter`$(R_j, e, \hat{f}_e^{new}, \hat{f}_e^{old})$
10 **else**
11 **if** $j == j_t$ **then** $\hat{f}_e^{new} = \hat{f}_e^{old}$ = `updateSketch`(M_{j_t}, e, w)
12 **if** $\hat{f}_e^{new} \le 2^{R_j.G.m}$ **then break**
13 $\langle e', \hat{f}_{e'}^{new}, \hat{f}_{e'}^{old}\rangle$ = `insertFilter`$(R_j, e, \hat{f}_e^{new}, \hat{f}_e^{old})$
14 **if** $e' == e$ **then break**
15 **if** $e' \neq$ **NIL** **then** // swap out kicked element e'
16 $j_t' = \min\big(\ell, \arg\max_{j'} \{\wedge_{1\le i\le j'} h_i(e') \neq 0\}\big)$
17 **if** $j_t' == j$ **then** `updateSketch`$(M_{j_t'}, e', \hat{f}_{e'}^{new} - \hat{f}_{e'}^{old})$
18 $\hat{L}_0 = \hat{L}_0 - 2^j \cdot (1 - 2 \cdot \mathbf{1}_{j+1\le j_t'}) \cdot g(\hat{f}_{e'}^{new})$
19 $j_b^{new} = j$ // track new bottom layer
20 $\hat{L}_0$ += $\mathbf{1}_{j_b^{new}\le j_t} \cdot 2^{j_b^{new}} \cdot g(\hat{f}_e^{new}) - \mathbf{1}_{j_b^{old}\le j_t} \cdot 2^{j_b^{old}} \cdot g(\hat{f}_e^{new} - w)$

the tuple (e, w) arrival, and it is 1 if $j_b^{new} \le j_t$. The latter is denoted by $\mathbf{1}_{j_b^{old}\le j_t} \cdot 2^{j_b^{old}} \cdot g(\hat{f}_e^{new} - w)$, where $\mathbf{1}_{j_b^{old}\le j_t}$ indicates whether e exists in at least one of the prefilters before the tuple arrival.

Update Moment by e'. The pseudocode to update the frequency moment $\hat{L}_0$ by the kicked element e' is given in Lines 15–18 of Algorithm 3. At Line 15, if the element e' returned at Line 13 is not **NIL**, a valid element e' is kicked out at Line 27 of Algorithm 1. If the topmost sampled layer j_t' of e' computed at Line 16 equals the current layer j, we swap out the element e' to the sketch on layer j at Line 17. Then, at Line 18, as the element e' is kicked out from the prefilter on the jth layer, we subtract its moment contribution $2^j \cdot (1 - 2 \cdot \mathbf{1}_{j+1\le j_t'}) \cdot g(\hat{f}_{e'}^{new})$ on layer j by OME expressed in Eq. (9).

D DATASET AND MATRICS

D.1 Metrics

The metrics used are as follows:

- **Average Absolute Error (AAE).** $AAE = \frac{1}{r}\sum_{1\le i\le r} |\hat{X}_i - X_i|$, where X_i and $\hat{X}_i$ are the actual and estimated value, respectively. The r is the number of distinct elements. We use AAE to evaluate the accuracy of per-element frequency estimation.
- **Average Relative Error (ARE).** $ARE = \frac{1}{r}\sum_{1\le i\le r} |\frac{\hat{X}_i - X_i}{X_i}|$. We use AAE to evaluate the accuracy of per-element frequency estimation and the frequency estimation of estimated heavy hitters.
- **F1 Score (F1).** $F1 = \frac{2*Precision\ Rate*Recall\ Rate}{Precision\ Rate+Recall\ Rate}$. We use the F1 score to evaluate the accuracy of heavy hitter detection.
- **Relative Error (RE).** $RE = \frac{|True-Estimated|}{True}$, where $True$ and $Estimated$ are the true and estimated values, respectively. We use RE to evaluate the accuracy of moment estimation.
- **Weighted Mean Relative Error (WMRE).** The $WMRE$ is calculated as $\big(\sum_{i=1}^{\mathbf{f}} |Y_i - \hat{Y}_i|\big) \big/ \big(\sum_{i=1}^{\mathbf{f}} \frac{Y_i+\hat{Y}_i}{2}\big)$, where $\mathbf{f}$ denotes the maximum element frequency, Y_i and $\hat{Y}_i$ represent the actual and estimated number of elements with frequency i, respectively.
- **Throughput.** Million of operations per second (Mops).

E ADDITIONAL EXPERIMENTS

E.1 Parameter Settings for *Maxkicks* of the KP-CF Prefilter

In this subsection, we measure the effects of the parameter $Maxkicks$, i.e., the maximum number of kicks for KP-CF, based on the CAIDA traces dataset. We set the relative threshold ε defined in Eq. (2) to 2^{-14}, and use AAE, ARE, F1 Score, as well as Throughput to evaluate the effect of $Maxkicks$. We show the results in Fig. 14.

As shown in Fig. 14, the AAE, ARE and F1 Score show little difference when $MaxKicks$ reaches 8, whereas the throughput decreases as $MaxKicks$ increases. So we set $MaxKicks$ to 8 in the following.

E.2 Parameter Settings for the Number of Hierarchical Layers of Universal Sketch

In Section 9.4, we present our strategy to choose the optimal number of hierarchical layers. The strategy suggests that, ideally, the universal sketch should contain $\lceil \log_2(N/k) \rceil + 1$ layers, where N represents the number of distinct elements in the data stream, and k is the prefilter size of each layer of the universal sketch. In this subsection, we perform experiments to determine the optimal number of hierarchical layers in practice, ranging from $\lceil \log_2(N/k) \rceil - 1$ to $\lceil \log_2(N/k) \rceil + 3$ (i.e., $\lceil \log_2(N/k) \rceil + 1 \pm 2$) for the CAIDA traces, IMDB, Yelp reviews, and synthetic Zipf distribution datasets.

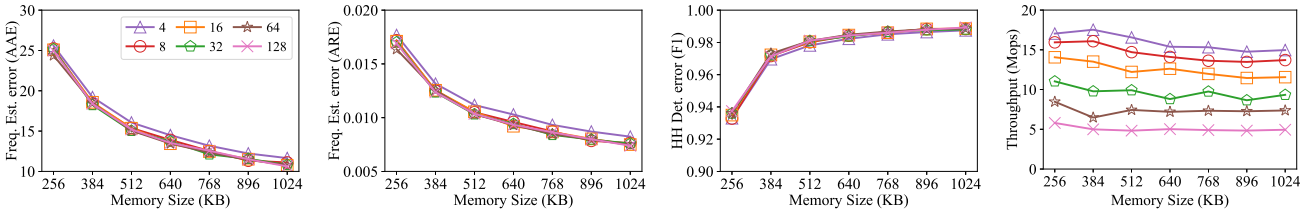
Parameter Settings. The settings for evaluating the optimal number of hierarchical layers for the four datasets are shown in Table 4. In the # Layers column, the value of $\lceil \log_2(N/k) \rceil + 1$ is highlighted in bold, such as **9** for the CAIDA traces dataset.

Table 4: Settings of evaluations on # hierarchical layers

Dataset	Memory	$D:I$	# Layers
CAIDA traces	From 768KB to 2560KB	0.5	7, 8, **9**, 10, 11
IMDB	From 768KB to 2560KB	0.5	9, 10, **11**, 12, 13
Yelp reviews	From 5M to 50M	0.5	12, 13, **14**, 15, 16
Zipf dist.	2560KB	From 0.1 to 0.9	10, 11, **12**, 13, 14

The explanation of the parameter settings is provided below.

- For the CAIDA traces dataset, we fix the $D:I$ ratio to 0.5, and set the memory allocation to range from 256KB to 2560KB. Given that the dataset contains 1.1M distinct elements as described in Section 9.1, the value of $\lceil \log_2(N/k) \rceil$ is calculated to be 8. Therefore, we select the optimal number of layers $\ell + 1$ from $\{7, 8, \mathbf{9}, 10, 11\}$.
- For the IMDB dataset, we fix the $D:I$ ratio to 0.5, and set the memory allocation to range from 256KB to 2560KB. Given that the dataset contains 3.1M distinct elements as described in Section 9.1,



(a) AAE, CAIDA dataset (b) ARE, CAIDA dataset (c) F1 Score, CAIDA dataset (d) Throughput, CAIDA dataset

Figure 14: Effects of the parameter *MaxKicks* on heavy hitter detection

the value of $\lceil \log_2(N/k) \rceil$ is calculated to be 10. Therefore, we select the optimal number of layers $\ell + 1$ from $\{9, 10, \mathbf{11}, 12, 13\}$.

- For the Yelp reviews dataset, we fix the $D:I$ ratio to 0.5, and set the memory allocation to range from 5MB to 50MB. Since the dataset contains 22.0M distinct elements as described in Section 9.1, the value of $\lceil \log_2(N/k) \rceil$ is calculated to be 13. Therefore, we select the optimal number of layers $\ell + 1$ from $\{12, 13, \mathbf{14}, 15, 16\}$.
- For the Zipf distribution dataset, we fix the memory allocation to 2560KB, and set the $D:I$ ratio to range from 0.1 to 0.9. We select the optimal number of hierarchical layers $\ell + 1$ from $\{10, 11, \mathbf{12}, 13, 14\}$. The reason is that, the dataset contains 4.9M distinct elements as described in Section 9.1, and the value of $\lceil \log_2(N/k) \rceil$ is calculated to be **11** by assuming $D:I = 0.5$.

Results for the Three Real-World Datasets with a Fixed $D:I$ Ratio (Figs. 15a–15i). The 0th-order moment (i.e., the number of distinct elements) estimation results for the CAIDA traces dataset, IMDB dataset, and Yelp reviews dataset with a fixed $D:I$ ratio are shown in Figs. 15a–15i. It is observed that the optimal number of hierarchical layers for the CAIDA traces dataset, IMDB dataset, and Yelp reviews dataset are 9, 10, and 15, respectively. Furthermore, once the number of hierarchical layers reaches the optimal value, the difference in the estimation becomes smaller than when the number of hierarchical layers is fewer than the optimal value. For example, as shown in Fig. 15a, when the number of layers is set to 10 or 11, the RE of the moment estimation is slightly higher than when the number of layers is set to 9. However, when the number of layers is set to 7 or 8, the RE of moment estimation is significantly higher than when the number of layers is set to 9, 10, or 11.

Results for the Synthetic Dataset with Varying $D:I$ Ratios (Figs. 15j–15o). The results for the synthetic Zipf distribution dataset (in both the shuffled and sorted patterns as described in Section 9.1) with varying $D:I$ ratios are shown in Figs. 15a–15i. The results show that, when the number of layers is set to 12, the RE of the 0th-order moment for the Zipf distribution dataset in both patterns is very close to that when the number of layers is set to 13 or 14. Therefore, we select 12 as the optimal number of layers.

E.3 Performance of Distribution Reconstruction

We use the distribution fitter MoM+CtJ proposed in [89] to reconstruct the frequency distribution. MoM+CtJ combines the method of moments (MoM) with the cut-then-rejoin strategy (CtJ), which cuts off the extra-large values from moments before applying MoM, as these values may cause poor convergence. We compare with MoM.

We use the CAIDA dataset and set the $D:I$ ratio to 0.5, then employ the online estimation of the moments from the 0th- to 7th-order for distribution reconstruction. We utilize the Weighted Mean Relative Error (WMRE) as a metric to evaluate the reconstruction.

The curve fitness and the WMRE of the MoM and MoM+CtJ are shown in Fig. 16. As shown in Fig. 16(a), MoM has a poor fitting performance, while MoM+CtJ fits the curve much more closely, especially for the high-frequency values. The WMRE of MoM shown in Fig. 16(b) is 0.090. Since we use the 0th- to 7th-order moments to reconstruct the per-element frequency distribution, the large values of the high-order moments may cause poor convergence of MoM. In contrast, by using the cut-then-rejoin strategy to remove the large values before applying the MoM, the MoM+CtJ can converge to a better result, and its WMRE shown in Fig. 16(b) is 0.004.

F MATHEMATICAL ANALYSIS

In this section, we present the mathematical analyses of our solutions. We first analyze the expectation and variance of our RAC. Then, we derive the L1 and L2 error bounds for RA-CS and RA-SuCS. Finally, for RAS, we provide the lower bound of the filter size and the error bound for heavy hitter detection.

F.1 Mathematical Analysis of RAC

In this subsection, we present the expected value of RAC to prove its unbiased nature. Then we present the variance of our RAC, and the coefficient of variation (i.e., $\frac{variance}{expectation^2}$).

Suppose an RAC C with an exponent α of length $\mathcal{L}_\alpha$, and a coefficient β of length $\mathcal{L}_\beta$. Denote by $T(a,b)$ the random variable that represents the amount of total update value needed from the start of counting ($C.\alpha = 0, C.\beta = 0$) until RAC C reaches the state when $C.\alpha$ equals a and $C.\beta$ equals b. We also assume a uniform increment or decrement size θ.

Theorem 2. *Suppose that increments to the RAC C are uniform and given by θ. Then, for the random variable $T(a,b)$ defined above we have:*

$$E[T(a,b)] = (b + 2^{\mathcal{L}_\beta}) \cdot 2^a - 2^{\mathcal{L}_\beta}$$

Proof. Recall that for a geometric random variable $G(p)$, we have $E[G(p)] = \frac{1}{p}$ and $Var[G(p)] = \frac{1-p}{p^2}$. For each a, denote by $W(a)$ the random variable that represents the total updates required for one increment of the RAC C when $C.\alpha = a$. The probability of incrementing the RAC C in a single trial is denoted as $p_a = \frac{\theta}{2^a}$. Note that the maximum value of $C.\alpha$ is $2^{\mathcal{L}_\alpha} - 1$. Therefore, the number of trials before the increment of RAC C is $G(p_a)$, and since

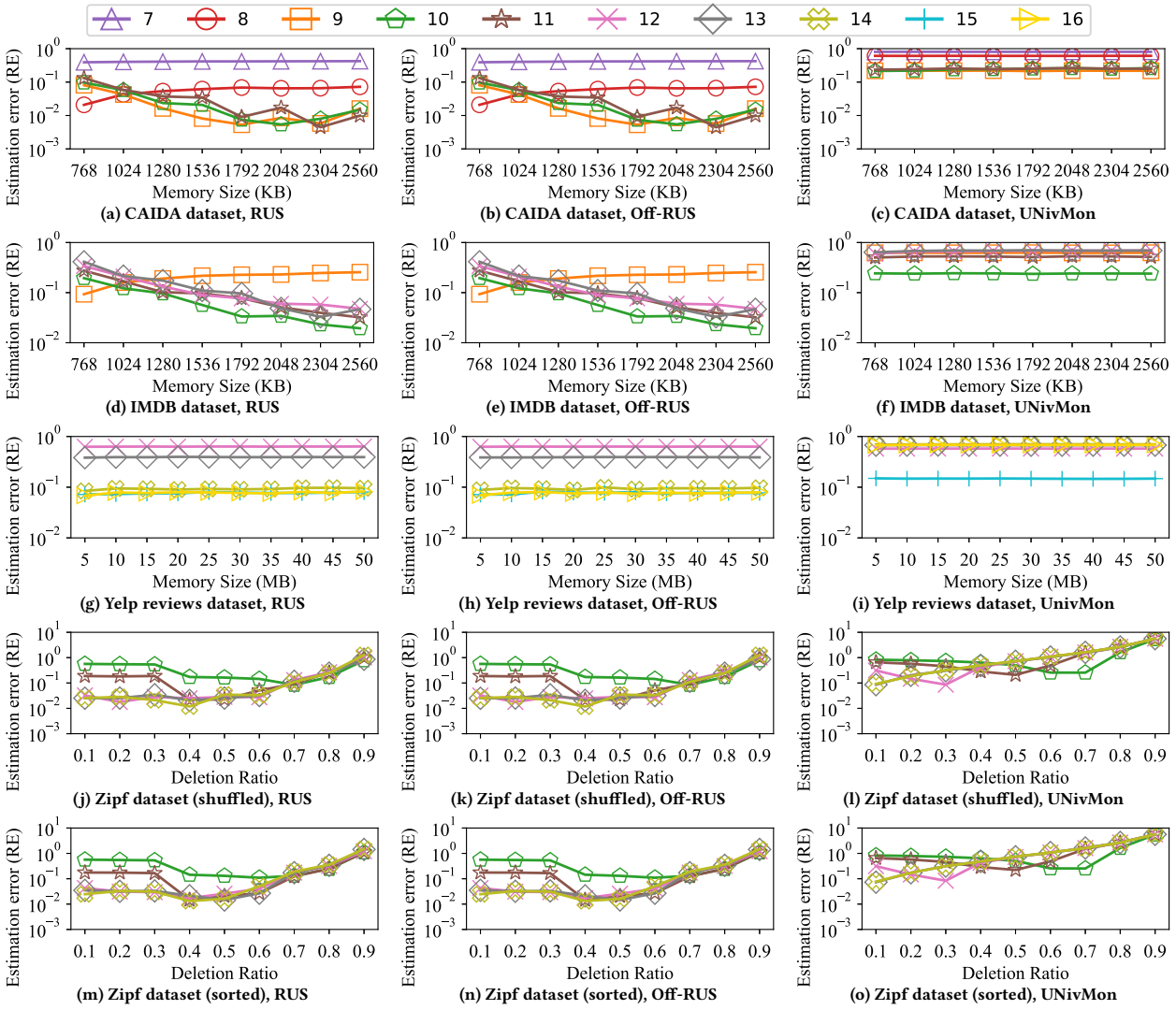
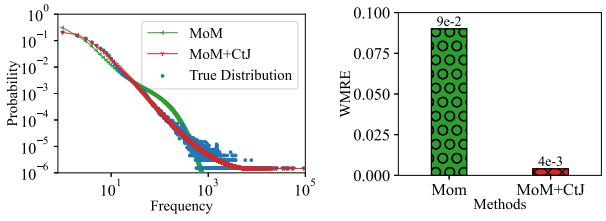


Figure 15: Effects of the number of hierarchical layers on the 0th-order moment (number of distinct elements N) estimation in the four datasets



(a) Curve fitness **(b) Accuracy**

Figure 16: Results of frequency distribution reconstruction

each trial corresponds to θ, we have

$$W(a) = \theta G(p_a)$$

Time from the beginning of counting can be divided in $2^{\mathcal{L}_\alpha}$ intervals, each corresponding to an exponent $\alpha = 0, 1, ..., a$, where $a < 2^{\mathcal{L}_\alpha}$. In the case of $\alpha = 0$, RAC C is incremented $q_0 = 2^{\mathcal{L}_\beta}$ times, for exponent $\alpha = 1, 2, ..., a \leq 2^{\mathcal{L}_\alpha} - 2$, we have $q_a = 2^{\mathcal{L}_\beta}$ times. For exponent $\alpha = 2^{\mathcal{L}_\alpha} - 1$, we have $q_a = 2^{\mathcal{L}_\beta} - 1$. This is because if we increment $2^{\mathcal{L}_\beta}$ times, the coefficient will overflow, and the exponent will increase 1 to correct this overflow, however, the maximum value of exponent α is $2^{\mathcal{L}_\alpha} - 1$. If $a = 2^{\mathcal{L}_\alpha}$, it overflows.

Denote by $Q(a)$ the total increments in each of these $2^{\mathcal{L}_\alpha}$ intervals. Now for any $\alpha = 0, 1, ..., a$, we have

$$Q(a) = \sum_{j=1}^{q_a} W_j(a)$$

where $W_j(a)$ is the set of independent identically distributed (i.i.d.) random variables, with distribution given by $W(a)$.

Consequently,

$$E[Q(a)] = \sum_{j=1}^{q_a} E[W_j(a)] = q_a \cdot \theta \frac{1}{p_a} = q_a \cdot 2^a$$

where $a \in \{0, 1, \ldots, 2^{\mathcal{L}_\alpha} - 1\}$, and $p_a = \frac{\theta}{2^a}$.
Now $T(a, b)$ is expressed as:

$$T(a, b) = \sum_{\alpha=0}^{a} Q(\alpha)$$

For a is 0, the expected value of $T(a, b)$ is

$$\begin{aligned} E[T(a,b)] &= E[\sum_{j=1}^{b} W_j(0)] \\ &= \sum_{j=1}^{b} E[\theta G(p_0)] \\ &= b \end{aligned}$$

The theorem holds when $a = 0$.
When $a \in 1, 2, \ldots, 2^{\mathcal{L}_\alpha} - 2$, the expected value of $T(a, b)$ is

$$\begin{aligned} E[T(a,b)] &= E[Q(0)] + \sum_{\alpha=1}^{a-1} E[Q(\alpha)] + b \cdot E[W(a)] \\ &= 2^{\mathcal{L}_\beta} + 2^{a+\mathcal{L}_\beta} - 2^{\mathcal{L}_\beta+1} + b \cdot 2^a \\ &= (b + 2^{\mathcal{L}_\beta}) \cdot 2^a - 2^{\mathcal{L}_\beta} \end{aligned}$$

The theorem holds when $a \in 1, 2, \ldots, 2^{\mathcal{L}_\alpha} - 2$.
When $a = 2^{\mathcal{L}_\alpha} - 1$, the expected value of $T(a, b)$ is

$$\begin{aligned} E[T(a,b)] &= E[Q(0)] + \sum_{\alpha=1}^{2^{\mathcal{L}_\alpha}-2} E[Q(\alpha)] + b \cdot E[W(a)] \\ &= 2^{\mathcal{L}_\beta} + 2^{2^{\mathcal{L}_\alpha}+\mathcal{L}_\beta-1} - 2^{\mathcal{L}_\beta+1} + b \cdot 2^{2^{\mathcal{L}_\alpha}-1} \\ &= (b + 2^{\mathcal{L}_\beta}) \cdot 2^a - 2^{\mathcal{L}_\beta} \end{aligned}$$

The theorem holds when $a = 2^{\mathcal{L}_\alpha} - 1$. □

Corollary 3. *The estimation of RAC C is unbiased.*

Proof. As stated in Theorem 2, the total update value equals the expectation of the estimation of RAC C. □

Theorem 4. *Suppose that increments to the RAC C are uniform and are given by θ. Then, for the random variable $T(a, b)$ defined above we have:*

$$\begin{aligned} Var[T(a,b)] \le \frac{1}{3}\left(2^{2^{\mathcal{L}_\alpha}} - 2\right) 2^{\mathcal{L}_\beta} \left(-3\theta + 2^{2^{\mathcal{L}_\alpha}} + 2\right) \\ -2^{2^{\mathcal{L}_\alpha}-2}\left(2^{2^{\mathcal{L}_\alpha}} - 2\theta\right) \end{aligned}$$

Proof. Since $Q(a) = \sum_{j=1}^{q_a} W_j(a)$, we have

$$\begin{aligned} Var[Q(a)] &= \sum_{j=1}^{q_a} Var[W_j(a)] \\ &= \sum_{j=1}^{q_a} Var[\theta G(p_a)] \\ &= q_a \cdot 2^a \cdot 2^a \cdot (1 - p_a) \end{aligned}$$

The variance of $T(a, b)$ can be expressed as follows:

$$\begin{aligned} Var(T(a,b)) &= \sum_{\alpha=0}^{a} Var[Q(a)] \\ &\le Var[Q(0)] + \sum_{a=1}^{2^{\mathcal{L}_\alpha}-2} Var[Q(a)] + Var[Q(\mathcal{L}_\alpha - 1)] \end{aligned}$$

For $Var[Q(0)]$, as the exponent part is 0, the increment is deterministic, so the variance is 0.

For $\sum_{\alpha=1}^{\mathcal{L}_\alpha-2} Var[Q(\alpha)]$, we have:

$$\begin{aligned} \sum_{\alpha=1}^{2^{\mathcal{L}_\alpha}-2} Var[Q(\alpha)] &= \sum_{\alpha=1}^{2^{\mathcal{L}_\alpha}-2} q_\alpha \cdot 2^{2\alpha}(1 - p_\alpha) \\ &= \sum_{\alpha=1}^{2^{\mathcal{L}_\alpha}-2} 2^{\mathcal{L}_\beta} \cdot 2^{2\alpha}(1 - \frac{\theta}{2^\alpha}) \\ &= \frac{1}{3}2^{2^{\mathcal{L}_\alpha+1}+\mathcal{L}_\beta-2} - \frac{1}{3}2^{\mathcal{L}_\beta+2} - \theta 2^{\mathcal{L}_\beta+2^{\mathcal{L}_\alpha}-1} + \theta 2^{\mathcal{L}_\beta+1} \end{aligned}$$

For $Var[Q(2^{\mathcal{L}_\alpha} - 1)]$, we have:

$$\begin{aligned} Var[Q(2^{\mathcal{L}_\alpha} - 1)] &= Var[\sum_{j=1}^{q_{2^{\mathcal{L}_\alpha}-1}} W_j(2^{\mathcal{L}_\alpha} - 1)] \\ &= (2^{\mathcal{L}_\beta} - 1) \cdot (2^{2^{\mathcal{L}_\alpha+1}-2} - \theta 2^{2^{\mathcal{L}_\alpha}-1}) \end{aligned}$$

Overall, the variance is bounded by:

$$\begin{aligned} Var[T(a,b)] &\le \sum_{a=0}^{\mathcal{L}_\alpha-1} Var[Q(a)] \\ &= \frac{1}{3}\left(2^{2^{\mathcal{L}_\alpha}} - 2\right) 2^{\mathcal{L}_\beta} \left(-3\theta + 2^{2^{\mathcal{L}_\alpha}} + 2\right) \\ &\quad -2^{2^{\mathcal{L}_\alpha}-2}\left(2^{2^{\mathcal{L}_\alpha}} - 2\theta\right) \end{aligned}$$

□

The following corollary characterizes the asymptotic behavior of the coefficient of variation $\delta(T(A, m))$.

Corollary 5. *For RAC C, $\delta(T(a, b)) \approx \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}}$.*

Proof.

$$\delta(T(a,b)) = \sqrt{\frac{Var[T(a,b)]}{(E[T(a,b)])^2}}$$

And we have

$$\begin{aligned} &\frac{Var[T(a,b)]}{(E[T(a,b)])^2} \\ &= \frac{\left(2^{2^{\mathcal{L}_\alpha}} - 2\right) 2^{\mathcal{L}_\beta+2}\left(-3\theta + 2^{2^{\mathcal{L}_\alpha}} + 2\right) - 3\ 2^{2^{\mathcal{L}_\alpha}}\left(2^{2^{\mathcal{L}_\alpha}} - 2\theta\right)}{3\left(2^{2^{\mathcal{L}_\alpha}} - \left(2^{2^{\mathcal{L}_\alpha}} - 1\right) 2^{\mathcal{L}_\beta+1}\right)^2} \\ &\approx \frac{2^{\mathcal{L}_\beta+2} - 3}{3\left(2^{\mathcal{L}_\beta+1} - 1\right)^2} \\ &\approx \frac{2^{-\mathcal{L}_\beta}}{3} \end{aligned} \tag{12}$$

where the approx symbol follows from that the $2^{2^{\mathcal{L}_\alpha}}$ is larger than the constant term in the addition process. □

F.2 Mathematical Analysis of RA-CS

In this subsection, we first prove that the estimation of RA-CS (i.e., CountSketch combined with our RAC) is unbiased. Then, we provide the L1 and L2 error bounds for RA-CS.

Theorem 6. *The RA-CS provides an unbiased element frequency estimation.*

Proof. We treat the value of RAC C as the true value C_r plus the estimation error C_ϵ, such that $C = C_r + C_\epsilon$. For estimating the frequency of an element e, in each row of the CountSketch, we have

$$\hat{f_e} = f_e + \sum_{e':e' \neq e} f_{e'} g(e) g(e') Y_{e'},$$

where

$$Y_{e'} = \begin{cases} 1, & if \ \ h(e') = h(e) \\ 0, & otherwise \end{cases}$$

So, for the frequency estimation $\hat{f_e}$ of the element e, we have

$$\begin{aligned} E[\hat{f_e}] &= E[C] \\ &= E[C_r] + E[C_\epsilon] \\ &= f_e + E[\sum_{e':e' \neq e} f_{e'} g(e) g(e') Y_{e'}] + E[C_\epsilon] \end{aligned}$$

Since the CountSketch and RAC C are unbiased, we have $E[C_r] = 0$ and $E[\sum_{e':e' \neq e} f_{e'} g(e) g(e') Y_{e'}] = 0$. Consequently, $E[\hat{f_e}] = f_e$, and the theorem holds. □

Theorem 7 (L1 error bound for RA-CS with general depth). *For RA-CS with depth $d = O(\log(1/\delta))$, and width b, the frequency estimation error $|f_e - \hat{f_e}|$ of an element e is at most $\frac{\epsilon}{b} F_1$ with probability $1 - \delta$.*

Proof. For the L1 bound in one row, we have

$$\begin{aligned} E[|f_e - \hat{f_e}|] &= E[|\sum_{e' \neq e} Y_{e'} \cdot f_{e'} \cdot g_j(e) \cdot g_j(e') + C_\epsilon|] \\ &\leq E[|\sum_{e' \neq e} Y_{e'} \cdot f_{e'} \cdot g_j(e) \cdot g_j(e')|] + E[|C_\epsilon|] \\ &= E[|\sum_{e' \neq e} Y_{e'} \cdot f_{e'} \cdot g_j(e) \cdot g_j(e')|] + E[|C - C_r|] \end{aligned}$$

For $E[|\sum_{e' \neq e} Y_{e'} \cdot f_{e'} \cdot g_j(e) \cdot g_j(e')|]$, it is bound by

$$E[|\sum_{e' \neq e} Y_{e'} \cdot f_{e'} \cdot g_j(e) \cdot g_j(e')|] \leq \frac{F_1}{b}$$

For $E[|C - C_r|]$, since the RAC C is unbiased (i.e., $C_r = E[C]$), we have

$$\begin{aligned} E[|C - C_r|] &\leq \sqrt{E[(C - C_r)^2]} = \sqrt{Var[C]} \\ &= \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3} \cdot (E[C])^2} \\ &= C_r \cdot \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}} \end{aligned}$$

Thus,

$$E[|f_e - \hat{f_e}|] \leq \frac{F_1}{b} + C_r \cdot \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}}$$

where the C_r can be seen as the expectation of the frequency in one RAC, which is $\frac{F_1}{b}$.

Hence,

$$\begin{aligned} E[|f_e - \hat{f_e}|] &\leq \frac{F_1}{b} + \frac{F_1}{b} \cdot \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}} \\ &= \frac{F_1}{b} \cdot (1 + \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}}) \end{aligned}$$

Note that the b equals twice that in the vanilla CountSketch.

Applying Markov's inequality, we deduce

$$P[|f_e - \hat{f_e}| \geq \frac{\epsilon}{b} F_1] \leq \frac{1}{\epsilon} \cdot (1 + \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})$$

Setting ϵ to 3, we have

$$P[|f_e - \hat{f_e}| \geq \frac{3}{b} F_1] \leq \frac{1}{3} \cdot (1 + \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})$$

$$P[|f_e - \hat{f_e}| \leq \frac{3}{b} F_1] \geq 1 - \frac{1}{3} \cdot (1 + \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})$$

Combining the Chernoff bounds, Theorem 7 follows.

□

Theorem 7 suggests using a value of b that is logarithmic in relation to the desired failure probability. However, practitioners rarely use more than a small constant number of rows, such as 3, 4, or 5. Recently, a study [34] proved that a CountSketch with a depth of 3 satisfies a similar error bound:

Theorem 8 (L1 error bound for RA-CS with a depth of 3). *For an RA-CS with a depth of 3, and a width of b, the frequency estimation error $|f_e - \hat{f_e}|$ of an element e is at most $\frac{\sqrt{3}}{b}(1 + \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}}) F_1$.*

Proof. Combining Corollary 5 with the study [34], the theorem holds. □

Theorem 9 (L2 error bound for RA-CS with general depth). *For an RA-CS with depth $d = O(\log(1/\delta))$, and width b, the frequency estimation error $|f_e - \hat{f_e}|$ of an element e is at most $\frac{\epsilon}{\sqrt{b}} F_2$ with probability $1 - \delta$.*

Proof. Given the computed variance of RAC C based on the expectation $E[C] = C_r$, we have the conditional variance $Var[C|C_r]$, and thus the variance of RAC C can be computed as

$$Var[C] = E[Var[C|C_r]] + Var[E[C|C_r]]$$

where $Var[C|C_r] = Var[T(a,b)] \approx \frac{2^{-\mathcal{L}_\beta}}{3} \cdot (E[T(a,b)])^2 < \frac{2^{-\mathcal{L}_\beta}}{3} \cdot \frac{{F_2}^2}{b}$, and $Var[E[C|C_r]]$ is $Var[C_r] \leq \frac{{F_2}^2}{b}$.

Then, we have

$$\begin{aligned} Var[C] &\leq E[\frac{2^{-\mathcal{L}_\beta}}{3} \cdot (E[T(a,b)])^2] + \frac{{F_2}^2}{b} \\ &= \frac{2^{-\mathcal{L}_\beta}}{3} E[(E[T(a,b)])^2] + \frac{{F_2}^2}{b} \\ &\leq \frac{2^{-\mathcal{L}_\beta}}{3} \frac{{F_2}^2}{b} + \frac{{F_2}^2}{b} \\ &= \frac{3 + 2^{-\mathcal{L}_\beta}}{3} \frac{{F_2}^2}{b} \end{aligned}$$

Note that the width b is twice that in the vanilla CountSketch.

By the Chebyshev's inequality, we have

$$P[|f_e - \hat{f_e}| \geq \frac{\epsilon}{\sqrt{b}} F_2] \leq \frac{3 + 2^{-\mathcal{L}_\beta}}{3} \cdot \frac{1}{\epsilon^2}$$

Setting ϵ to 2, we have

$$P[|f_e - \hat{f_e}| \geq \frac{2}{\sqrt{b}} F_2] \leq \frac{3 + 2^{-\mathcal{L}_\beta}}{3} \cdot \frac{1}{4}$$

$$P[|f_e - \hat{f_e}| \leq \frac{2}{\sqrt{b}} F_2] \geq 1 - \frac{3 + 2^{-\mathcal{L}_\beta}}{3} \cdot \frac{1}{4}$$

By the Chernoff bounds, Theorem 9 follows. □

Theorem 10 (L2 error bound for RA-CS with a depth of 3). *For an RAS with depth 3 and width b, the frequency estimation error $|f_e - \hat{f}_e|$ of an element e is at most $\sqrt{\frac{3+2^{-\mathcal{L}_\beta}}{3}}\frac{F_2}{\sqrt{b}}$.*

Proof. Combining Corollary 5 with the study [34], the theorem holds. □

F.3 Mathematical Analysis of RA-SuCS

In this subsection, we first prove that the SuCS and RA-SuCS (i.e., SuCS combined with our RAC) provide an unbiased estimation. Then, we derive the L1 and L2 error bound of the RA-SuCS.

Theorem 11. *The SuCS provides an unbiased element frequency estimation.*

Proof. Assume f_e^i represents the frequency of the element e that is hashed to its i-th counter, where $i < 4$ and the counter is randomly chosen. For the i-th counter, we have

$$\hat{f}_e^i = \sum\nolimits_{e':e'\neq e} f_{e'} g(e) g(e') Y_{e'}^i, \tag{13}$$

where

$$Y_{e'}^i = \begin{cases} 1, & if \;\; h^i(e') = h^i(e) \\ 0, & otherwise \end{cases}$$

Therefore, for the estimation of f_e^i, we have

$$E[\hat{f}_e^i] = f_e^i + E\big[\sum\nolimits_{e':e'\neq e} f_{e'} g(e) g(e') Y_{e'}^i\big]$$

Since $E[g(e')] = 0$, we have $E[\hat{f}_e^i] = f_e^i$

So, the estimation of SuCS is unbiased, and the theorem holds. □

Theorem 12. *The RA-SuCS provides an unbiased element frequency estimation.*

Proof. Combining Corollary 3 and Theorem 11, the theorem holds. □

Theorem 13 (L1 error bound for RA-SuCS). *For an RA-SuCS with width b, the frequency estimation error $|f_e - \hat{f}_e|$ of an element e is at most $\frac{\epsilon}{b}F_1$ with probability $1 - \frac{4}{\epsilon}\cdot(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})$, where $\mathcal{L}_\beta$ is the length of the coefficient part in an RAC.*

Proof. For the L1 bound of one of e's hashed counters, we have

$$\begin{aligned} E[|f_e^i - \hat{f}_e^i|] &= E[|\sum\nolimits_{e'\neq e} Y_{e'}^i \cdot f_{e'}^i \cdot g^i(e)\cdot g^i(e') + C_\epsilon|] \\ &\le E[|\sum\nolimits_{e'\neq e} Y_{e'}^i \cdot f_{e'}^i \cdot g^i(e)\cdot g^i(e')|] + E[|C_\epsilon|] \\ &= E[|\sum\nolimits_{e'\neq e} Y_{e'}^i \cdot f_{e'}^i \cdot g^i(e)\cdot g^i(e')|] + E[|C - C_r|] \end{aligned}$$

For $E[|\sum_{e'\neq e} Y_{e'}^i \cdot f_{e'}^i \cdot g^i(e)\cdot g^i(e')|]$, we have

$$E[|\sum\nolimits_{e'\neq e} Y_{e'}^i \cdot f_{e'}^i \cdot g^i(e)\cdot g^i(e')|] \le \frac{F_1}{b}$$

For $E[|C - C_r|]$, since the RAC C is unbiased (i.e., $C_r = E[C]$), we have

$$\begin{aligned} E[|C - C_r|] &\le \sqrt{E[(C-C_r)^2]} = \sqrt{Var[C]} \\ &= \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}\cdot(E[C])^2} \\ &= C_r \cdot \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}} \end{aligned}$$

Therefore,

$$E[|f_e^i - \hat{f}_e^i|] \le \frac{F_1}{b} + C_r \cdot \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}}$$

where the C_r can be seen as the expectation of the frequency in one RAC, equivalent to $\frac{F_1}{b}$.

Therefore,

$$\begin{aligned} E[|f_e^i - \hat{f}_e^i|] &\le \frac{F_1}{b} + \frac{F_1}{b}\cdot\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}} \\ &= \frac{F_1}{b}\cdot(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}}) \end{aligned}$$

Additionally,

$$E[|f_e - \hat{f}_e|] \le \frac{4F_1}{b}\cdot(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})$$

Note that the b is eight times that in the vanilla CountSketch and double that in SuCS.

Combine Markov's inequality, we have

$$P[|f_e - \hat{f}_e| \ge \frac{\epsilon}{b}F_1] \le \frac{4}{\epsilon}\cdot(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})$$

Setting ϵ to 12, we have

$$P[|f_e - \hat{f}_e| \ge \frac{12}{b}F_1] \le \frac{1}{3}\cdot(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})$$

$$P[|f_e - \hat{f}_e| \le \frac{12}{b}F_1] \ge 1 - \frac{1}{3}\cdot(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})$$

Combining the Chernoff bounds, Theorem 13 follows.

□

Theorem 14 (L2 error bound for RA-SuCS). *For an RA-SuCS with width b, the frequency estimation error $|f_e - \hat{f}_e|$ of an element e is at most $\frac{\epsilon}{\sqrt{b}}F_2$ with probability $1 - \frac{48+2^{-\mathcal{L}_\beta}}{3}\frac{1}{\epsilon^2}$, where $\mathcal{L}_\beta$ is the length of the coefficient part in an RAC.*

Proof. We have computed the variance of the RAC C, based on the expectation $E[C] = C_r$. So we have the conditional variance $Var[C|C_r]$, and the variance of RAC C can be computed as

$$Var[C] = E[Var[C|C_r]] + Var[E[C|C_r]]$$

where $Var[C|C_r] = Var[T(a,b)] \approx \frac{2^{-\mathcal{L}_\beta}}{3}\cdot(E[T(a,b)])^2 < \frac{2^{-\mathcal{L}_\beta}}{3}\cdot\frac{{F_2}^2}{b}$, and $Var[E[C|C_r]]$ is $Var[C_r] \le \frac{16{F_2}^2}{b}$.

Then, we have

$$\begin{aligned} Var[C] &\le E[\frac{2^{-\mathcal{L}_\beta}}{3} \cdot (E[T(a,b)])^2] + \frac{16{F_2}^2}{b} \\ &= \frac{2^{-\mathcal{L}_\beta}}{3} E[(E[T(a,b)])^2] + \frac{16{F_2}^2}{b} \\ &\le \frac{2^{-\mathcal{L}_\beta}}{3} \frac{{F_2}^2}{b} + \frac{16{F_2}^2}{b} \\ &= \frac{48 + 2^{-\mathcal{L}_\beta}}{3} \frac{{F_2}^2}{b} \end{aligned}$$

Note that the width b is 8 times as large as that of the CountSketch, and twice as large as that of the SuCS.

By the Chebyshev's inequality, we have

$$P[|f_e - \hat{f}_e| \ge \frac{\epsilon}{\sqrt{b}} F_2] \le \frac{48 + 2^{-\mathcal{L}_\beta}}{3} \cdot \frac{1}{\epsilon^2}$$

Setting ϵ to 8, we have

$$P[|f_e - \hat{f}_e| \ge \frac{8}{\sqrt{b}} F_2] \le \frac{48 + 2^{-\mathcal{L}_\beta}}{3} \cdot \frac{1}{64}$$

$$P[|f_e - \hat{f}_e| \le \frac{8}{\sqrt{b}} F_2] \ge 1 - \frac{48 + 2^{-\mathcal{L}_\beta}}{3} \cdot \frac{1}{64}$$

By the Chernoff bounds, Theorem 14 follows. □

F.4 Mathematical Analysis of RAS

In this subsection, we first discuss the lower bound on the size of the prefilter, followed by the error bound of the frequency estimation for heavy hitters.

Lower Bound on Filter Size. Previously, SpaceSaving$^{\pm}$ (SS$^{\pm}$) [90] established a lower bound $L = \frac{\zeta}{\varepsilon}$ on its size to track all heavy hitters (HHs) in the *bounded deletion model*. However, the lower bound is determined under the assumption of an evenly distributed insertion pattern. In this pattern, the number of insertions of each element is equal; therefore, no HHs exist. In this section, we further derive the lower bound. In this section, we derive a new lower bound by considering a data stream that follows the Zipf Law, characterized by heavy hitters, a distribution common in real-world data streams.

The lower bound when the data streams follow a Zipf distribution is given in Theorem 1, with the proof outlined below.

Proof. The probability mass function for the Zipf distribution is shown in Eq. (14), where η represents the exponent that characterizes the skewness of the distribution, and N is the number of distinct elements.

$$f(i;\eta,N) = 1 \,/\, \left(i^\eta \sum_{j=1}^{N} j^{-\eta}\right) \tag{14}$$

After all I insertions and D deletions, the frequencies of heavy hitters are at least $\frac{\varepsilon}{\zeta} I$, since $\varepsilon(I - D) = \frac{\varepsilon}{\zeta} I$. Thus, the prefilter must retain all the elements with a frequency greater than $\frac{\varepsilon}{\zeta} I$ before any deletion occurs. Otherwise, if such an element is not retained before deletion occurs, it becomes an untracked heavy hitter afterward. Therefore, the size lower bound L is the maximum value of k that satisfies $f(k;\eta,N) \ge \frac{\varepsilon}{\zeta}$. The expression of L is defined as

$$L = \max k, \text{ s.t. } 1 \le k \le N \ \wedge \ f(k;\eta,N) \ge \frac{\varepsilon}{\zeta}.$$

Combining Eq. (14), the value of k can be determined as

$$k \le \left\lfloor \sqrt[\eta]{\frac{\zeta/\varepsilon}{\sum_{i=1}^{N}(1/i)^\eta}} \right\rfloor. \tag{15}$$

□

The derivative of $\sqrt[\eta]{\frac{\zeta/\varepsilon}{\sum_{i=1}^{N}(1/i)^\eta}}$ with respect to η is negative, i.e., $\frac{\partial}{\partial\eta} \sqrt[\eta]{\frac{\zeta/\varepsilon}{\sum_{i=1}^{N}(1/i)^\eta}} < 0$. As the parameter η describes the skewness, this indicates that a Zipf distribution with less skewness requires a larger lower bound on the size of the prefilter. Consequently, the configuration of k suitable for a low-skew Zipf distribution also applies to a high-skew Zipf distribution. Therefore, we derive the corollary for the lower bound k when $\eta = 1$ (i.e., low skew Zipf distribution) to fit more scenarios.

Corollary 15. *For a Zipf distribution with skewness parameter η no less than* 1*, the lower bound on the prefilter size can be established to $\frac{\zeta}{12\varepsilon}$.*

Proof. Since the derivative of $\sqrt[\eta]{\frac{\zeta/\varepsilon}{\sum_{i=1}^{N}(1/i)^\eta}}$ shown in Eq. (15) with respect to η is negative, a Zipf distribution with large skewness requires a lower bound on the size of the prefilter that is even lower. Therefore, if the lower bound is applicable when η equals 1, it also applies when η is larger than 1.

When η is 1, which indicative of a low-skew Zipf distribution [31, 55], the size lower bound L can be expressed as

$$k = \frac{\zeta/\varepsilon}{\sum_{i=1}^{N}(1/i)} \approx \frac{\zeta/\varepsilon}{ln(N+1)+\gamma}, \tag{16}$$

where $\gamma \approx 0.5772$ is the Euler's constant.

By combining Theorem 1 with the assumption that N typically exceeds 2^{16} in real-world datasets, we set the size of the filter to $\frac{\zeta/\varepsilon}{12}$, which is only $\frac{1}{12}$ of that of the SS$^{\pm}$. This value is validated to yield excellent results in the evaluation presented in **Section 9**. □

Estimation Error of RAS. From Theorem 1, we transform the ε-heavy hitters problem to the top-k heavy hitter problem, where k is set to $\frac{\zeta}{\varepsilon \sum_{i=1}^{N}(1/i)}$. Let f_e^k denote the frequency of the kth most frequent element. We then give the L1 and L2 error bound to ensure that all the elements with frequency at least $(1-\varepsilon)f_e^k$ are maintained.

Theorem 16 (L1 bound of the heavy hitter problem when using RA-CS with a depth of 3). *If d is set to* 3*, and $b \ge \frac{2\sqrt{3}(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})F_1}{\epsilon f_e^k}$, then all elements with a frequency no less than $(1-\epsilon)f_e^k$ among the top-$\frac{\zeta}{\varepsilon \sum_{i=1}^{N}(1/i)}$ element are preserved.*

Proof. By Theorem 8, the estimation for the frequency of all elements is within an additive factor of $\frac{\sqrt{3}}{b}(1 + \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})F_1$ of the actual element frequency. Thus for two elements whose true frequency differs by more than $2 \cdot \frac{\sqrt{3}}{b}(1 + \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})F_1$, the estimation can correctly identify the more frequent element. By setting $2 \cdot \frac{\sqrt{3}}{b}(1 + \sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})F_1 \le \epsilon f_e^k$, we ensure that the only elements

that can replace the true most frequent elements in the estimated top-$\frac{\zeta}{\varepsilon \sum_{i=1}^{N}(1/i)}$ are those with true frequency at least $(1-\epsilon)f_e^k$.

$$2 \cdot \frac{\sqrt{3}}{b}(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})F_1 \leq \epsilon f_e^k$$

$$b \geq \frac{2\sqrt{3}(1+\sqrt{\frac{2^{-\mathcal{L}_\beta}}{3}})F_1}{\epsilon f_e^k}$$

□

Theorem 17 (L1 bound of the heavy hitter problem when using RA-CS with general depth). *If $b \geq \frac{6F_1}{\epsilon f_e^k}$, then the element with frequency no less than $(1-\epsilon)f_e^k$ among the top-$\frac{\zeta}{\epsilon \sum_{i=1}^{N}(1/i)}$ elements are preserved with probability $1-\delta$ from the RAS with parameters $d = O(\log(1/\delta))$.*

Proof. Similar to Theorem 16, we combine Theorem 7 and set the error smaller than half of ϵf_e^k, the theorem holds. □

Theorem 18 (L1 bound of the heavy hitter problem when using RA-SuCS). *If $b \geq \frac{24F_1}{\epsilon f_e^k}$, then the element with frequency no less than $(1-\epsilon)f_e^k$ among the top-$\frac{\zeta}{\epsilon \sum_{i=1}^{N}(1/i)}$ elements are preserved with probability $1-\delta$ from the RAS with parameters $d = O(\log(1/\delta))$.*

Proof. Similar to Theorem 16, we combine Theorem 13 and set the error smaller than half of ϵf_e^k, the theorem holds. □

Theorem 19 (L2 bound of the heavy hitter problem when using RA-CS with a depth of 3). *If d is set to 3, and $b \geq \frac{4}{3}{F_2}^2 \frac{(3+2^{-\mathcal{L}_\beta})^2}{(\epsilon f_e^k)^2}$, then the elements with frequency no less than $(1-\epsilon)f_e^k$ among the top-$\frac{\zeta}{\epsilon \sum_{i=1}^{N}(1/i)}$ element are preserved.*

Proof. The proof is similar to that of Theorem 16. Combined with Theorem 10, when $b \geq \frac{4}{3}{F_2}^2 \frac{(3+2^{-\mathcal{L}_\beta})^2}{(\epsilon f_e^k)^2}$, the estimation error is smaller than half of $(1-\epsilon)f_e^k$, so the theorem holds. □

Theorem 20 (L2 bound of the heavy hitter problem when using RA-CS with general depth). *If $b \geq \frac{16{F_2}^2}{\epsilon^2 {f_e^k}^2}$, then the element with frequency no less than $(1-\epsilon)f_e^k$ among the top-$\frac{\zeta}{\epsilon \sum_{i=1}^{N}(1/i)}$ element are preserved with probability $1-\delta$ from the RAS with parameters $d = O(\log(1/\delta))$.*

Proof. The proof is similar to that of Theorem 16. Combined with Theorem 9, when $b \geq \frac{16{F_2}^2}{\epsilon^2 {f_e^k}^2}$, the estimation error is smaller than half of $(1-\epsilon)f_e^k$, so the theorem holds. □

Theorem 21 (L2 bound of the heavy hitter problem when using RA-SuCS). *If $b \geq \frac{256{F_2}^2}{\epsilon^2 {f_e^k}^2}$, then the element with frequency no less than $(1-\epsilon)f_e^k$ amont the top-$\frac{\zeta}{\epsilon \sum_{i=1}^{N}(1/i)}$ element are preserved with probability $1-\delta$ from the RAS with parameters $d = O(\log(1/\delta))$.*

Proof. The proof is similar to that of Theorem 16. Combined with Theorem 14, when $b \geq \frac{256{F_2}^2}{\epsilon^2 {f_e^k}^2}$, the estimation error is smaller than half of $(1-\epsilon)f_e^k$, so the theorem holds. □