A MATHEMATICAL ANALYSIS

In this section, we will present the error bounds and expected precision of the Bubble Sketch for top-k items.Consistent with the previous discussion, our analysis is based on the assumption that the probability distribution of the items follows a Zipf distribution.Typically, the skewness parameter α is not less than 1, and the larger the value of α, the more accurate the estimation of the top-k items.For simplicity, we assume that $\alpha = 1$.

In this section, the specific structure of the Bubble Sketch we derived is as follows: There are w buckets in *Array1* and *Array2*, each with a size of 128 bits, divided into 5 entries. Among them, *entry5* contains a 32-bit full ID and a 32-bit counter, while entries 1, 2, 3, and 4 each store an 8-bit fingerprint and counters of sizes 4 bits, 4 bits, 8 bits, and 16 bits, respectively.

A.1 Average absolute error (AAE)

A.1.1 Overestimation Error Bound. During the insertion process, the main overestimation comes from fingerprint collusion. However, when a item is stored in *entry5*, since it stores the complete 32-bit ID, fingerprint collusion are almost impossible; and since we always return the maximum value in the current bucket, this value should most of the time be stored in *entry5*. Therefore, our Bubble Sketch has a lower error upper bound, which is explained in detail below.

A.1.2 Lemma 1. For projects A and B, their respective counts are a and b ($a > b$). Inserting them one by one in a random order, and considering the time for each insertion as *1*, let T represent the expected time during the insertion process when the count of B is higher than that of A. Then, we have:

$$T \leq \frac{\sqrt{\frac{a}{b}}}{(\sqrt{\frac{a}{b}} - 1)^2} \tag{1}$$

Proof: In practical situations, where a and b are relatively large, we can simply assume that the arrival of each item is independent. That is, each arriving item has a probability of $\frac{a}{a+b}$ of being A and a probability of $\frac{b}{a+b}$ of being B. Consider the probability that the count of B is greater than that of A when the s^{th} item arrives is represented as t_s, then:

$$t_s = \sum_{i=\frac{s}{2}}^{s} C_s^i \cdot (\frac{b}{a+b})^i \cdot (\frac{a}{a+b})^{s-i} \tag{2}$$

$$\leq \sum_{i=\frac{s}{2}}^{s} C_s^i \cdot (\frac{b}{a+b})^{\frac{s}{2}} \cdot (\frac{a}{a+b})^{\frac{s}{2}} \tag{3}$$

$$= 2^{s-1} \cdot (\frac{b}{a+b})^{\frac{s}{2}} \cdot (\frac{a}{a+b})^{\frac{s}{2}} \tag{4}$$

$$= \frac{1}{2} \cdot (\frac{2\sqrt{ab}}{a+b})^s \tag{5}$$

In that case, we arrive at the following conclusion or formula:

$$T = \sum_{i=1}^{a+b} t_i \tag{6}$$

$$\leq \sum_{i=1}^{a+b} \frac{1}{2} \cdot (\frac{2\sqrt{ab}}{a+b})^i \tag{7}$$

$$= \frac{1}{2} \cdot \frac{2\sqrt{ab}}{a+b} \cdot \sum_{i=1}^{a+b} (\frac{2\sqrt{ab}}{a+b})^{i-1} \tag{8}$$

$$\leq \frac{\sqrt{ab}}{a+b} \cdot \frac{1}{1 - \frac{2\sqrt{ab}}{a+b}} \tag{9}$$

$$= \frac{\sqrt{\frac{a}{b}}}{(\sqrt{\frac{a}{b}} - 1)^2} \tag{10}$$

Now considering a top-k item f_j, regarding its overestimation bound, we can draw the following conclusion:

Theorem A.1. *For overestimation errors, we have:*

$$E(\hat{n_j} - n_j) \leq \frac{n}{1024wn_j + 2n} \cdot ((P_j n_j + (1 - P_j)\frac{\theta}{(\theta - 1)^2}) \tag{11}$$

where $P_j = (\frac{2w-1}{2w})^{\frac{n_1}{n_j}}$, $\theta = \sqrt{\frac{n_j}{\frac{n_j}{2} + \frac{n}{512w}}}$

Proof: To estimate the expectation of $\hat{n_j} - n_j$, let's denote n_t as the value of the second-largest item in the bucket where n_j is located. As mentioned before, for an overestimation to occur, n_j must not be in *entry5*, so we are particularly interested in the size of n_k. We denote T as the expected time when n_t is greater than n_j.T is larger when n_t is close to n_j, but this scenario is relatively less probable; T is smaller when n_t is far from n_j, which can be derived from the *Lemma 1*. Given our dataset follows a Zipf distribution, let n_j be the l_j^{th} largest item among all items, and n_t be the t_j^{th} largest. Then, $l_j = \frac{n_1}{n_j}$, and $l_j < k, l_j < l_t$. We consider a coefficient $\eta > 1$, using θ times l_j as the delimiter.For t less than ηl_j, we simply assume that the time for $n_t > n_j$ does not exceed $\frac{1}{2}n_j$; for t greater than ηl_j, according to Lemma 1, we know that the time for $n_t > n_j$ does not exceed $\frac{\theta}{(\theta-1)^2}$, where $\theta$is the result of taking $a = n_j$ and b as the estimated value of n_t in Lemma 1,and among these, only half of the time is spent inserting n_j. .For simplicity, let's first set η to 2.When n_j is not in the highest position, the probability of n_j being overestimated is $\frac{\frac{n}{512w}}{n_j + \frac{n}{512w}}$.The probability of t being between l_j and ηl_j is $(\frac{2w-1}{2w})^{l_j}$.So that, we have:

$$\begin{aligned} E(\hat{n_j} - n_j) &\leq \frac{\frac{n}{512w}}{n_j + \frac{n}{512w}} (\frac{1}{2}(\frac{2w-1}{2w})^{l_j} n_j + (1 - (\frac{2w-1}{2w})^{l_j})\frac{\theta}{(\theta - 1)^2}) \\ &= \frac{n}{1024wn_j + 2n} \cdot ((P_j n_j + (1 - P_j)\frac{\theta}{(\theta - 1)^2}) \end{aligned}$$

where $P_j = (\frac{2w-1}{2w})^{\frac{n_1}{n_j}}$, $\theta = \sqrt{\frac{n_j}{\frac{n_j}{2}+\frac{n}{512w}}}$.

A.1.3 Underestimation Error Bound.

Theorem A.2. *For underestimation errors, we have:*

$$E(n_j - \hat{n_j}) \leq 2(1 - Q_j)^4 \cdot Q_j^{-1} \tag{12}$$

where $Q_j = \frac{512wn_j+n}{512wn_j+256n}$

proof: We note that underestimation can only occur when n_j is located at entry1.We calculate Q_j, the conditional probability that among all items mapped to the bucket where n_j is located, there are items with the same 8-bit fingerprint as:

$$\begin{aligned} Q_j &= \frac{n_j + \frac{n}{512w}}{nj + \frac{n}{2}} \\ &= \frac{512wn_j + n}{512wn_j + 256n} \end{aligned}$$

If n_j does not start in entry1, given n_j's absolute dominance over other items, it is almost impossible for n_j to be demoted to entry1. However, if n_j is in entry1, then when entry1 is 0 and n_j occurs consecutively, n_j will move up. The probability of entry1 being 0 is $\frac{1}{2}$, and the expected occurrence of consecutive n_j is Q_j^{-1}, so we have:

$$E(n_j - \hat{n_j}) \leq 2(1 - Q_j)^4 \cdot Q_j^{-1}$$

where $Q_j = \frac{512wn_j+n}{512wn_j+256n}$.

A.2 Precision

In this section, we first provide an estimate of the expected precision of the Bubble Sketch under the most ideal conditions (insertion of only the top-k items). Then, we present an estimate of the expected precision of the Bubble Sketch considering the insertion order of all items. Finally, we estimate the impact of false negatives due to the limited space in the buckets, resulting in a practical estimate of the expected precision of the Bubble Sketch.

A.2.1 Theorem 1. Let p represent the precision of the Bubble Sketch (Precision is defined as the proportion of accurately identified top-k items out of the total k items reported), assuming only top-k items are inserted. Then, we have:

$$E(p) \geq \frac{4w - k - 1}{2w} + \frac{e^{-\frac{2}{w}}}{k}(\frac{1 - e^{-\frac{2k}{w}}}{1 - e^{-\frac{2}{w}}}) - \frac{e^{-\frac{1}{w}}}{k}(2 - \frac{1}{w(1 - e^{-\frac{1}{w}})}) - \frac{e^{-\frac{k+1}{w}}}{w(1 - e^{-\frac{1}{w}})} \tag{13}$$

This result represents the error caused by the interactions among the top-k items.

Proof:Since each bucket can only retain one potential top-k items, we directly abstract the structure of the Bubble sketch as each bucket having only 1 entry. When inserting an item, if the candidate bucket in Array1 is full, it attempts to insert into the candidate bucket in Array2. If the candidate bucket in Array2 is also full, then the insertion fails (equivalent to a cuckoo hash without a kick action). Assuming that when the c-th item is inserted, there are s full buckets in Array1, Then, regarding s, we have:

$$\frac{w}{w} + \frac{w}{w-1} + ... + \frac{w}{w-s+1} \approx c \tag{14}$$

then we have:

$$E(s) \approx w(1 - e^{-\frac{c}{w}}) \tag{15}$$

Let $f(c)$ be the probability that the c-th item fails to insert on both attempts, then

$$\begin{aligned} f(c) &\le E(\frac{s}{w} \cdot \frac{c-s}{w}) \\ &\le \frac{E(s)}{w} \cdot \frac{c - E(s)}{w} \\ &= (1 - e^{-\frac{c}{w}}) \cdot (\frac{c}{w} - 1 + e^{-\frac{c}{w}}) \end{aligned}$$

so that:

$$\begin{aligned} E(p) &= 1 - \frac{1}{k}\sum_{c=1}^{k} f(c) \\ &\ge 1 - \frac{1}{k}\sum_{c=1}^{k}(1 - e^{-\frac{c}{w}}) \cdot (\frac{c}{w} - 1 + e^{-\frac{c}{w}}) \\ &= \frac{4w-k-1}{2w} + \frac{e^{-\frac{2}{w}}}{k}(\frac{1-e^{-\frac{2k}{w}}}{1-e^{-\frac{2}{w}}}) - \frac{e^{-\frac{1}{w}}}{k}(2 - \frac{1}{w(1-e^{-\frac{1}{w}})}) - \frac{e^{-\frac{k+1}{w}}}{w(1-e^{-\frac{1}{w}})} \end{aligned}$$

A.2.2 Theorem 2. Let p represent the precision of the bubble sketch, ignoring only the impact of estimation errors on items, we have:

$$E(p) \ge 1 - \frac{(k-1)(2k+1)}{24w^2} \tag{16}$$

proof: Consider the top-k items being inserted in order,let $f(c)$ be the probability that the c-th item fails to insert on both attempts, assuming there are s full buckets in Array1, then

$$\begin{aligned} f(c) &\le \frac{s}{w} \cdot \frac{c-1-s}{w} \\ &\le \frac{1}{4} \cdot \frac{(s-1)^2}{w^2} \end{aligned}$$

so that:

$$\begin{aligned} E(p) &= 1 - \frac{1}{k}\sum_{c=1}^{k} f(c) \\ &\geq 1 - \frac{1}{4k}\sum_{c=1}^{k}\frac{(s-1)^2}{w^2} \\ &= 1 - \frac{(k-1)(2k-1)}{24w^2} \end{aligned}$$

A.2.3 Theorem 3. Let p represent the precision of the Bubble Sketch, considering the error inherent in the sketch itself, we have:, we have:

$$E(p) \geq 1 - \frac{(k-1)(2k+1)}{24w^2} - P_e \tag{17}$$

where $P_e = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=k+1}^{2w}(\frac{1}{2w}(1-\frac{1}{2w})^{i-k-1}\cdot\frac{\frac{n}{512w}}{\frac{n_1}{j}-\frac{n_1}{i}})$

proof: For the j-th largest item, we only need to consider the probability that the second largest item in a bucket exceeds this item. Assuming that this second largest flow is the i-th largest, where i does not exceed $2w$, then the expected overestimation does not exceed $\frac{n}{512w}$.The difference between it and the i-th largest item is $\frac{n_1}{i} - \frac{n_1}{j}$.Then, according to the Minkowski inequality and considering the probability that it is precisely the i-th item, we obtain the probability for this case as $(\frac{1}{2w}(1-\frac{1}{2w})^{i-k-1}\cdot\frac{\frac{n}{512w}}{\frac{n_1}{j}-\frac{n_1}{i}})$ Then, the overall error probability is:

$$P_e \leq \frac{1}{k}\sum_{j=1}^{k}\sum_{i=k+1}^{2w}(\frac{1}{2w}(1-\frac{1}{2w})^{i-k-1}\cdot\frac{\frac{n}{512w}}{\frac{n_1}{j}-\frac{n_1}{i}})$$

According to the Theorem 2, we have:

$$E(p) \geq 1 - \frac{(k-1)(2k+1)}{24w^2} - P_e \tag{18}$$

where $P_e = \frac{1}{k}\sum_{j=1}^{k}\sum_{i=k+1}^{2w}(\frac{1}{2w}(1-\frac{1}{2w})^{i-k-1}\cdot\frac{\frac{n}{512w}}{\frac{n_1}{j}-\frac{n_1}{i}})$