SteadySketch: Finding Steady Flows in Data Streams (Appendix)

Anonymous Author(s)

February 12, 2023

A Error Bound of RollingSketch

In this part, we assume that our RollingSketch shares the same number of counters with strawman solution using 32 bits for each counter and both RollingSketch and strawman solution use only one hash function. We then provide an error bound for RollingSketch. In other words, we save 75% memory without sacrificing the performance of RollingSketch too much.

Lemma 1. *For a flow e and an arbitrary timestamp t, let g_0, g_1 be its frequency reported by RollingSketch, f be its frequency reported by strawman solution, then $(f - g_0)\%256 = 0, |g_1 - g_0| = 128$.*

Proof. This lemma is a simple property of rebirth. Note that rebirth happens only when $g_0 = 256$, and its value will change to 0 immediately, hence the two equality holds. □

Lemma 2. *For a flow e and p consecutive windows $t - p + 1, \cdots, t$, let $X_1, \cdots, X_p$ denotes the frequency of e in each window reported by strawman solution and*

$$M = \max\{X_i : 1 \leqslant i \leqslant p\}, m = \min\{X_i : 1 \leqslant i \leqslant p\}.$$

If $M - m < 128$, then RollingSketch will report the same variance as strawman solution.

Proof. We resort $X_1, \cdots, X_p$ as $x_1, \cdots, x_p$, s.t. $x_1 \leqslant \cdots \leqslant x_p$. Since $M - m = x_p - x_1 < 128$, there exists some integer m, s.t. $0 \leqslant x_1 - 128m \leqslant x_2 - 128m \leqslant \cdots \leqslant x_k \leqslant 127 < 128 \leqslant x_{k+1} - 128m \leqslant \cdots \leqslant x_p - 128m \leqslant 255$. This way of calculating variance keeps the sequence of number unchanged, so it can correctly reflect the variance of the statistics. Next we show that this value is actually the minimum among our calculation of variance. Note that $(x + 256m)\%256 = x\%256$, hence the only way for a smaller variance is

$$\begin{cases} y_1 \triangleq (x_1 + 128)\%256 = x_1 + 128, \\ \cdots \\ y_k \triangleq (x_k + 128)\%256 = x_k + 128, \\ y_{k+1} \triangleq (x_{k+1} + 128)\%256 = x_{k+1} - 128, \\ \cdots \\ y_p = \triangleq (x_p + 128)\%256 = x_p - 128. \end{cases}$$

If $k = 0$ or $k = n$, then all x_i will increment or decrement 128, which makes the variance unchanged; otherwise, define $\bar{x} = \frac{1}{p}(x_1 + \cdots + x_p)$, $u_i = x_i - \bar{x}$, and $s^2, \hat{s}^2$ be the variance of $x_1, \cdots, x_p; y_1, \cdots, y_p$. Then $u_1 + \cdots + u_p = 0$, and

$$s^2 = \frac{1}{p}\sum_{i=1}^{p}(x_i - \bar{x}) = \frac{1}{p}\sum_{i=1}^{p} u_i^2, \tag{1}$$

$$\begin{aligned}
\hat{s}^2 =& \frac{1}{p}\sum_{i=1}^{p}(y_i - \bar{y})^2 \\
=& \frac{1}{p}\left(\sum_{i=1}^{k}(u_i + 128)^2 + \sum_{j=k+1}^{p}(u_j - 128)^2\right) \\
& - \frac{1}{p^2}\left[128(2k-p)\right]^2 \\
=& \frac{1}{p}\left[\sum_{i=1}^{p} u_i^2 + 256\left(\sum_{i=1}^{k} u_i - \sum_{j=k+1}^{p} u_j\right) + 128^2 p\right. \\
& \left. -128^2\frac{(2k-p)^2}{p}\right] \\
=& \frac{1}{p}\left[\sum_{i=1}^{p} u_i^2 + 512(u_1 + \cdots + u_k) + \frac{128^2(4pk - 4k^2)}{p}\right].
\end{aligned} \tag{2}$$

Hence

$$\begin{aligned}
\hat{s}^2 - s^2 &= \frac{512}{p^2}\left[p(u_1 + \cdots + u_k) + 128(kp - k^2)\right] \\
&= \frac{512}{p^2}\left(kp\bar{u}^- + 128kp - 128k^2\right),
\end{aligned} \tag{3}$$

where $\bar{u}^-$ is defined as $\bar{u}^- = \frac{1}{k}(u_1 + \cdots + u_k)$. Since $x_p - x_1 < 128$, then $u_p - u_1 < 128$. Hence for $i = k+1, \cdots, p$, $u_i \leqslant u_p < 128 + u_1 \leqslant 128 + \bar{u}^-$. Finally, we get

$$\begin{aligned}
&\begin{cases} u_1 + \cdots + u_k \leqslant k\bar{u}^-, \\ u_{k+1} + \cdots + u_p \leqslant (p-k)(128 + \hat{u}^-) \end{cases} \\
\Rightarrow & 0 = \sum_{i=0}^{p} u_i \leqslant k\bar{u}^- + (p-k)(128 + \hat{u}^-) \\
\Rightarrow & \hat{u}^- \geq \frac{128(k-p)}{p}.
\end{aligned} \tag{4}$$

Applying this inequality, we get

$$\hat{s}^2 - s^2 \geqslant \frac{512}{p^2}\left[128k(k-p) + 128kp - 128k^2\right] = 0, \tag{5}$$

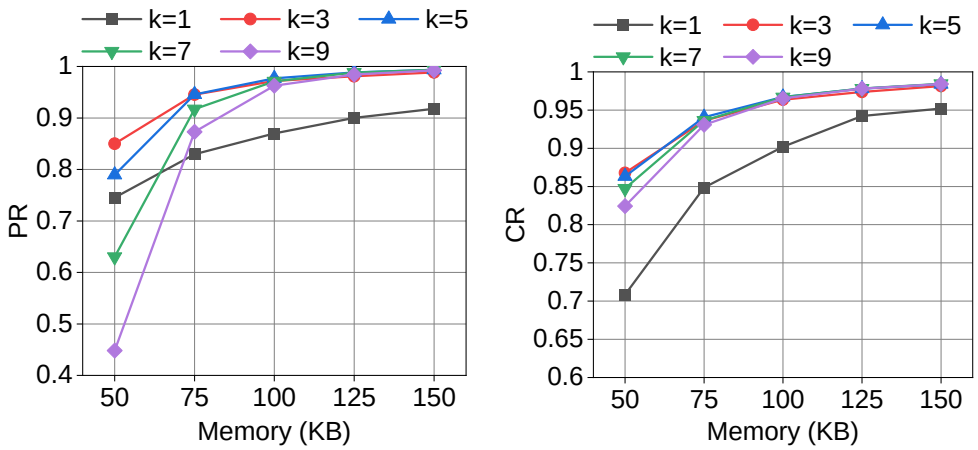
which shows that the variance reported by the RollingSketch is accurate. □

Theorem 3. *For a flow e, assume its frequency in each window is reported to be $X_1, X_2, \cdots$ by strawman solution. If RollingSketch wrongly reports $\langle e, t\rangle$ as steady flow but strawman solution does not, then $\exists t - p + 1 \leqslant i \neq j \leqslant t$, s.t. $X_i - X_j \geqslant 128$.*

Since steady flows shall occur similar times in each window, we conclude that these flows cannot be steady flows.

B Experiments on Parameter Settings

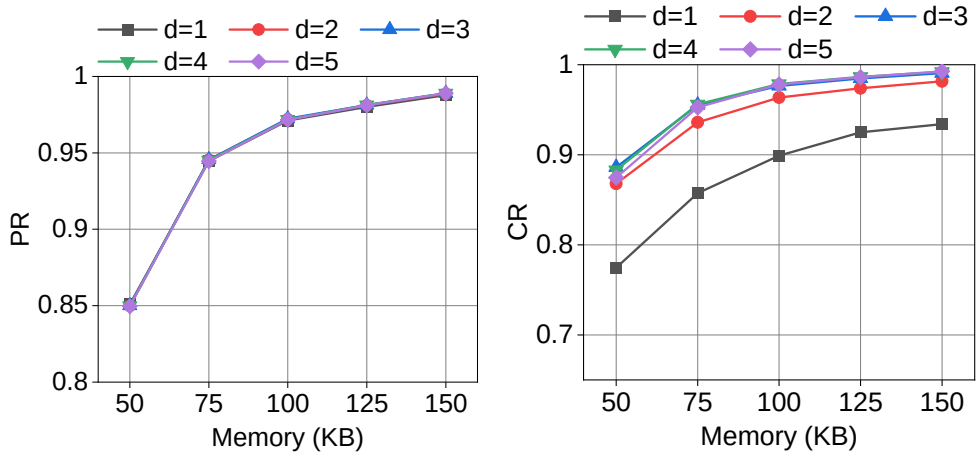
In this section, we measure the effects of some key parameters of SteadySketch, namely, the number of hash functions k in SteadyFilter, the ratio r of the memory size of SteadyFilter to the memory size of the whole SteadySketch, the number of hash functions d in GroupSketch, the the variance threshold H for the steady items, and the threshold p for time window period. We use the CAIDA dataset in these experiments, and PR and CR to evaluate the effects.



PR of k CR of k

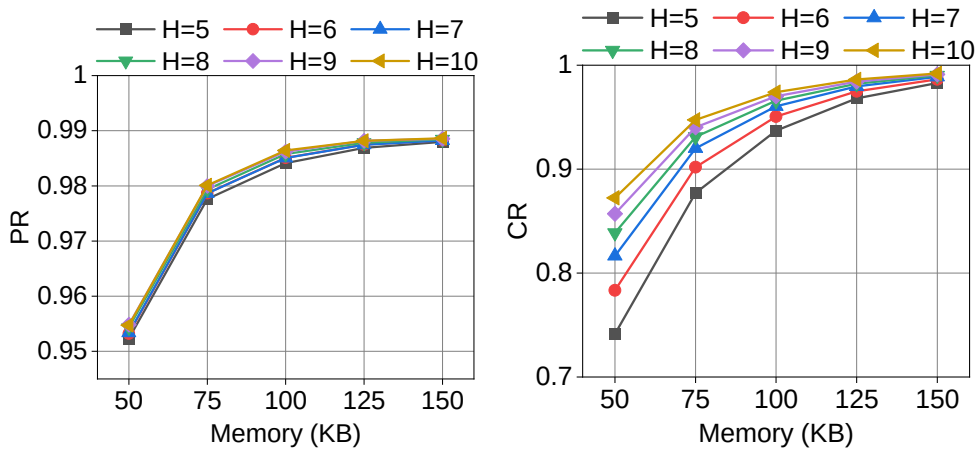
Figure 1: Effect of the parameter: k

1) Effect of k (Figure 1(a) - 1(b)): *The experimental results show that the best value for k is 3.* In this experiment, we fix the d to 2. Under the same memory, as k becomes larger, PR increases first and then decreases, while CR increases first and then does not change significantly. Thus, we set $k = 3$.
2) Effect of d (Figure 2(a) - 2(b)): *The experimental results show that the best value for d is 2.* In this experiment, we fix the k to 3. Under the same memory, as d becomes larger, CR increases first and then decreases, while PR does not change significantly. Thus, we set $d = 2$.
3) Effect of H (Figure 3(a) - 3(b)): *Our experimental results show that the PR and CR of different H values are close.* In this experiment, we fix the k to 3 and d to 2. As H becomes larger, CR gradually increases but gradually approaches as the memory



PR of d CR of d

Figure 2: Effect of the parameter: d



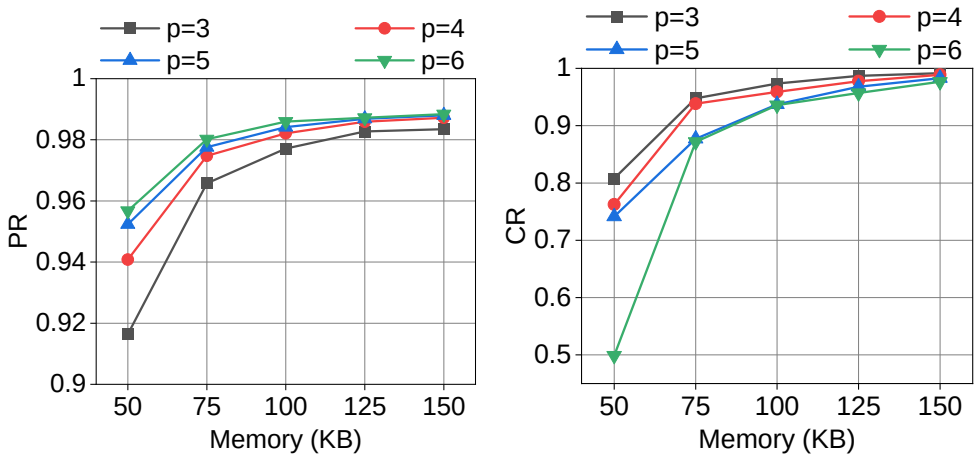
PR of H CR of H

Figure 3: Effect of the parameter: H

becomes larger, while PR does not change significantly. Since this value is as small as possible, we finally choose $H = 5$.

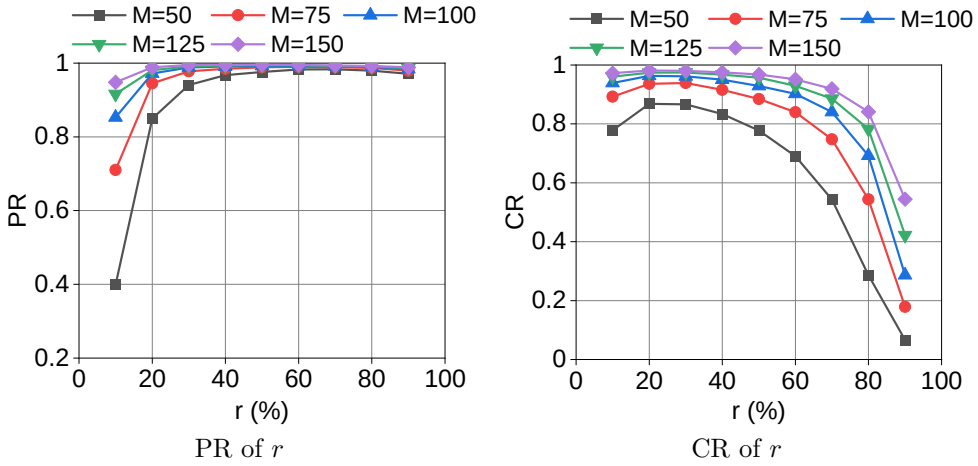
4) Effect of p (Figure 4(a) - 4(b)): *The experimental results show that the optimal value for p is 5.* In this experiment, we fix the k to 3, d to 2 and H to 5. As p becomes larger, PR gradually increases but gradually approaches as the memory becomes larger, while CR gradually decreases but gradually approaches as the memory becomes larger. Therefore, we set $p = 5$.

5) Effect of r (Figure 5(a) - 5(b)): *The experimental results show that the optimal value for r is 40%.* In this experiment, we fix the k to 3, d to 2, H to 5 and p to 5. M refers to the total memory of SteadySketch, which consists of the memory of SteadyFilter and RollingSketch. As r becomes larger, PR gradually increases, while CR



PR of p CR of p

Figure 4: Effect of the parameter: p



PR of r CR of r

Figure 5: Effect of the parameter: r

increases first and then decreases. Therefore, we choose $r = 20\%$ because it can trade off PR and CR well for different values of M.