Technical Report for "Identifying Hierarchical Super Spreaders in a Data Stream by Hot-Separated and Mergeable Sketch"

In this section, we analyze the estimation bias and standard deviation of the MOPS sketch. Since each row gives an independent estimated spread $\hat{n}_f^{(r)}$ of the flow f, which are aggregated by median operator, we start by analyzing the probabilistic distribution of $\hat{n}_f^{(r)}$ in the r-th row.

Preliminary of a Row. For a given flow f, all elements from other flows, totaling $n - n_f$, constitute noise that may be mapped to the d virtual estimators. Each noise element $\langle f', e\rangle$ with $f' \neq f$ has an equal probability of being mapped to any register in the d virtual estimators, due to the following independent sources of randomness:

(i) Random column selection via $h^{(r)}(f') \bmod w$,
(ii) Random estimator selection via $z^{(r)}(f') \bmod 2$,
(iii) Random row selection via $h^{(r)}(e) \bmod d$.

Let $n_{f|0}^{(r)}$ denote the number of flow-element pairs mapped to the *main* estimator for flow f in row r. The quantity $n_{f|0}^{(r)} - n_f$ thus represents the number of noise elements mapped to the main estimator, while $n_{f|1}^{(r)}$ denotes the number of noise elements mapped to the *alternative* estimator.

Under the conditions:

$$\begin{aligned} w &\gg 1 \quad \text{(large enough columns)}, \\ n_f &\ll n \quad \text{(small flow spread)}, \end{aligned}$$

the quantities $n_{f|0}^{(r)} - n_f$ and $n_{f|1}^{(r)}$ approximately follow the binomial distribution $\mathrm{Bino}(n - n_f, \frac{1}{2w})$. Specifically, for any integer $i \in [0, n - n_f]$,

$$\Pr\left\{n_{f|0}^{(r)} - n_f = i\right\} = \binom{n - n_f}{i}\left(\frac{1}{2w}\right)^i\left(1 - \frac{1}{2w}\right)^{n - n_f - i} \tag{1}$$

and for $j \in [0, n - n_f]$,

$$\Pr\left\{n_{f|1}^{(r)} = j\right\} = \binom{n - n_f}{j}\left(\frac{1}{2w}\right)^j\left(1 - \frac{1}{2w}\right)^{n - n_f - j} \tag{2}$$

The expected value and variance are:

$$\mathrm{E}\left(n_{f|0}^{(r)} - n_f\right) = \frac{1}{2w}(n - n_f) \tag{3}$$

$$\mathrm{E}\left(n_{f|1}^{(r)}\right) = \frac{1}{2w}(n - n_f) \tag{4}$$

$$\mathrm{Var}\left(n_{f|0}^{(r)} - n_f\right) = \frac{1}{2w}\left(1 - \frac{1}{2w}\right)(n - n_f) \tag{5}$$

$$\mathrm{Var}\left(n_{f|1}^{(r)}\right) = \frac{1}{2w}\left(1 - \frac{1}{2w}\right)(n - n_f) \tag{6}$$

In the following analysis, we consider the distributions of the following four random variables:

- The true number of noise elements $n_{f|0}^{(r)} - n_f$ mapped to the main estimator,
- The true number of noise elements $n_{f|1}^{(r)}$ mapped to the alternative estimator,
- The HLL-TC-based spread estimation $\hat{n}_{f|0}^{(r)}$ derived from the main estimator,
- The HLL-TC-based spread estimation $\hat{n}_{f|1}^{(r)}$ derived from the alternative estimator.

To characterize the estimation accuracy of $\hat{n}_{f|0}^{(r)}$ and $\hat{n}_{f|1}^{(r)}$, we invoke the following well-known result regarding the behavior of the HyperLogLog-TailCut (HLL-TC) estimator:

Theorem 1 (HyperLogLog-TailCut Error)**.** *Let n_m be the spread of a multiset that is sufficiently large, and let $\hat{n}_m$ be its estimation using HLL-TC with m registers. Then, the spread estimate $\hat{n}_m$ is asymptotically almost unbiased, i.e.,*

$$\frac{E(\hat{n}_m)}{n_m} = 1 + \delta_1(n_m) + o(1), \tag{7}$$

where $|\delta_1(n_m)| < 5 \times 10^{-5}$ as soon as $m \geq 16$.

Moreover, the standard error satisfies

$$\frac{\sqrt{\mathrm{Var}(\hat{n}_m)}}{n_m} = \frac{\gamma_m}{\sqrt{m}} + \delta_2(n_m) + o(1), \tag{8}$$

where $|\delta_2(n_m)| < 5 \times 10^{-4}$ for $m \geq 16$. The constant γ_m is bounded with $\gamma_m \to 1.04$ for HLL-TC when $m \geq 128$.

Estimation Bias of A Row. According to Theorem 1 and leveraging (1), we examine the conditional expectation of the main and alternative estimators. Conditioning on the event that $n_{f|0}^{(r)} - n_f = i$, we have:

$$\begin{aligned} \mathrm{E}\left(\hat{n}_{f|0}^{(r)} \mid n_{f|0}^{(r)} - n_f = i\right) &= (n_f + i)\left(1 + \delta_1(n_f + i) + o(1)\right) \\ &\approx n_f + i, \end{aligned} \tag{9}$$

where $\delta_1(\cdot)$ denotes the asymptotically vanishing bias term inherent in the HLL-TC estimator.

Similarly, conditioning on $n_{f|1}^{(r)} = j$, which accounts only for foreign flows, we obtain:

$$\mathrm{E}\left(\hat{n}_{f|1}^{(r)} \mid n_{f|1}^{(r)} = j\right) = j \cdot (1 + \delta_1(j) + o(1)) \approx j. \tag{10}$$

By the law of total expectation, the unconditional expectation of the main estimator is given by:

$$\begin{aligned}
\mathrm{E}\left(\hat{n}_{f|0}^{(r)}\right) &= \sum_{i=0}^{n-n_f} \mathrm{E}\left(\hat{n}_{f|0}^{(r)} \mid n_{f|0}^{(r)} - n_f = i\right) \cdot \Pr\left\{n_{f|0}^{(r)} - n_f = i\right\} \\
&\approx \sum_{i=0}^{n-n_f} (n_f + i) \cdot \binom{n-n_f}{i} \left(\frac{1}{2w}\right)^i \left(1 - \frac{1}{2w}\right)^{n-n_f-i} \\
&\approx n_f + \sum_{i=0}^{n-n_f} i \cdot \binom{n-n_f}{i} \left(\frac{1}{2w}\right)^i \left(1 - \frac{1}{2w}\right)^{n-n_f-i} \\
&= n_f + \frac{1}{2w}(n - n_f) \qquad (11)
\end{aligned}$$

where the last step uses the expectation of a Binomial random variable.

Likewise, the expected value of the alternative estimator becomes:

$$\begin{aligned}
\mathrm{E}\left(\hat{n}_{f|1}^{(r)}\right) &= \sum_{j=0}^{n-n_f} \mathrm{E}\left(\hat{n}_{f|1}^{(r)} \mid n_{f|1}^{(r)} = j\right) \cdot \Pr\left\{n_{f|1}^{(r)} = j\right\} \\
&\approx \sum_{j=0}^{n-n_f} j \cdot \binom{n-n_f}{j} \left(\frac{1}{2w}\right)^j \left(1 - \frac{1}{2w}\right)^{n-n_f-j} \\
&= \frac{1}{2w}(n - n_f) \qquad (12)
\end{aligned}$$

The final per-row estimator subtracts the alternative estimate from the main estimate to isolate the contribution of flow f:

$$\begin{aligned}
\mathrm{E}\left(\hat{n}_f^{(r)}\right) &= \mathrm{E}\left(\hat{n}_{f|0}^{(r)}\right) - \mathrm{E}\left(\hat{n}_{f|1}^{(r)}\right) \\
&\approx n_f + \frac{1}{2w}(n - n_f) - \frac{1}{2w}(n - n_f) \\
&= n_f. \qquad (13)
\end{aligned}$$

Therefore, the per-row estimator $\hat{n}_f^{(r)}$ is asymptotically unbiased for the true flow spread n_f.

Estimation Variance of A Row. Let flow f hash to bucket $c = h_r(\tilde{f})$ in row r, and let $z_r(\tilde{f}) \in \{0, 1\}$ indicate the choice of one of two sub-registers. Denote by $\hat{n}_{f|0}^{(r)}$ the spread estimate from the main estimator selected by $z_r(\tilde{f})$, and by $\hat{n}_{f|1}^{(r)}$ the estimate from the other alternative estimator. The per-row estimator is given by:

$$\hat{n}_f^{(r)} = \hat{n}_{f|0}^{(r)} - \hat{n}_{f|1}^{(r)}. \qquad (14)$$

Assuming that the two HLL-TC estimators are approximately independent due to contributions from disjoint sets of flows, the variance of the row estimator is:

$$\mathrm{Var}(\hat{n}_f^{(r)}) = \mathrm{Var}(\hat{n}_{f|0}^{(r)}) + \mathrm{Var}(\hat{n}_{f|1}^{(r)}), \qquad (15)$$

$$\mathrm{Var}(\hat{n}_{f|0}^{(r)}) = \mathrm{E}\left((\hat{n}_{f|0}^{(r)})^2\right) - \left(\mathrm{E}(\hat{n}_{f|0}^{(r)})\right)^2, \qquad (16)$$

$$\mathrm{Var}(\hat{n}_{f|1}^{(r)}) = \mathrm{E}\left((\hat{n}_{f|1}^{(r)})^2\right) - \left(\mathrm{E}(\hat{n}_{f|1}^{(r)})\right)^2. \qquad (17)$$

According to Theorem 1, the coefficient of variation under the condition $n_{f|0}^{(r)} = n_f + i$, for $i \in [0, n - n_f)$, is:

$$\begin{aligned}
&\frac{1}{n_f + i}\sqrt{\mathrm{Var}\left(\hat{n}_{f|0}^{(r)} \mid n_{f|0}^{(r)} = n_f + i\right)} \\
&= \frac{\gamma_m}{\sqrt{m}} + \delta_2(n_f + i) + o(1) \approx \frac{\gamma_m}{\sqrt{m}} \qquad (18)
\end{aligned}$$

and similarly for $n_{f|1}^{(r)}$ under condition $\hat{n}_{f|1}^{(r)} = j$:

$$\frac{1}{j}\sqrt{\mathrm{Var}\left(\hat{n}_{f|1}^{(r)} \mid n_{f|1}^{(r)} = j\right)} = \frac{\gamma_m}{\sqrt{m}} + \delta_2(j) + o(1) \approx \frac{\gamma_m}{\sqrt{m}} \qquad (19)$$

Thus, for $d \geq 128$, the conditional variances are:

$$\mathrm{Var}(\hat{n}_{f|0}^{(r)} \mid n_{f|0}^{(r)} = n_f + i) \approx \frac{\gamma_m^2}{m}(n_f + i)^2, \qquad (20)$$

$$\mathrm{Var}(\hat{n}_{f|1}^{(r)} \mid n_{f|1}^{(r)} = j) \approx \frac{\gamma_m^2}{m} j^2, \qquad (21)$$

where $\gamma_m \approx 1.04$ for HLL-TC.

From this, the conditional second moments follow:

$$\begin{aligned}
&\mathrm{E}\left(\left(\hat{n}_{f|0}^{(r)}\right)^2 \middle| n_{f|0}^{(r)} = n_f + i\right) \\
&= \mathrm{Var}\left(\hat{n}_{f|0}^{(r)} \mid n_{f|0}^{(r)} = n_f + i\right) + \left(\mathrm{E}\left(\hat{n}_{f|0}^{(r)} \mid n_{f|0}^{(r)} = n_f + i\right)\right)^2 \\
&\approx \left(\frac{\gamma_m^2}{m} + 1\right)(n_f + i)^2 \qquad (22)
\end{aligned}$$

Similarly, for the alternative estimator:

$$\begin{aligned}
&\mathrm{E}\left(\left(\hat{n}_{f|1}^{(r)}\right)^2 \middle| n_{f|1}^{(r)} = j\right) \\
&= \mathrm{Var}\left(\hat{n}_{f|1}^{(r)} \mid n_{f|1}^{(r)} = j\right) + \left(\mathrm{E}\left(\hat{n}_{f|1}^{(r)} \mid n_{f|1}^{(r)} = j\right)\right)^2 \\
&\approx \left(\frac{\gamma_m^2}{m} + 1\right)(j)^2 \qquad (23)
\end{aligned}$$

Applying the law of total expectation to the main estimator:

$$\begin{aligned}
&\mathrm{E}\left(\left(\hat{n}_{f|0}^{(r)}\right)^2\right) \qquad (24) \\
&= \sum_{i=0}^{n-n_f} \mathrm{E}\left(\left(\hat{n}_{f|0}^{(r)}\right)^2 \middle| n_{f|0}^{(r)} = n_f + i\right) \cdot \Pr\{n_{f|0}^{(r)} = n_f + i\} \\
&\approx \sum_{i=0}^{n-n_f} \left(\frac{\gamma_m^2}{m} + 1\right)(n_f + i)^2 \cdot \binom{n-n_f}{i}\left(\frac{1}{2w}\right)^i \left(1 - \frac{1}{2w}\right)^{n-n_f-i} \\
&= \left(\frac{\gamma_m^2}{m} + 1\right)\left((n_f)^2 + 2n_f \mathrm{E}(n_{f|0}^{(r)} - n_f) + \mathrm{E}((n_{f|0}^{(r)} - n_f)^2)\right) \\
&= \left(\frac{\gamma_m^2}{m} + 1\right) \\
&\cdot \left((n_f)^2 + 2n_f \mathrm{E}(n_{f|0}^{(r)} - n_f) + (\mathrm{E}(n_{f|0}^{(r)} - n_f))^2 + \mathrm{Var}(n_{f|0}^{(r)} - n_f)\right) \\
&= \left(\frac{\gamma_m^2}{m} + 1\right)\left((n_f + \frac{1}{2w}(n - n_f))^2 + \frac{1}{2w}\left(1 - \frac{1}{2w}\right)(n - n_f)\right) \qquad (25)
\end{aligned}$$

To simplify (25), we define two symbols A and B:

$$A = \mathrm{E}(n_{f|0}^{(r)}) = n_f + \frac{1}{2w}(n - n_f) \qquad (26)$$

$$B = \mathrm{Var}(n_{f|0}^{(r)} - n_f) = \frac{1}{2w}\left(1 - \frac{1}{2w}\right)(n - n_f) \qquad (27)$$

where:

- A is the expected number of elements mapped to the main estimator,
- B is the variance due to noise from other flows.

Applying the two symbols to (11) and (25), we have:

$$\left(\mathrm{E}\left(\hat{n}_{f|0}^{(r)}\right)\right)^2 = A^2, \tag{28}$$

$$\mathrm{E}\left(\hat{n}_{f|0}^{(r)} \cdot \hat{n}_{f|0}^{(r)}\right) = \left(\frac{\gamma_m^2}{m}+1\right)(A^2+B). \tag{29}$$

Then, we can substitute these into (16):

$$\begin{aligned}\mathrm{Var}\left(\hat{n}_{f|0}^{(r)}\right) &= \mathrm{E}\left(\hat{n}_{f|0}^{(r)} \cdot \hat{n}_{f|0}^{(r)}\right) - \left(\mathrm{E}\left(\hat{n}_{f|0}^{(r)}\right)\right)^2 &(30)\\ &= \left(\frac{\gamma_m^2}{m}+1\right)(A^2+B) - A^2 &(31)\\ &= \left(\frac{\gamma_m^2}{m}\right)A^2 + \left(\frac{\gamma_m^2}{m}+1\right)B. &(32)\end{aligned}$$

Similarly, for the alternative estimator:

$$\begin{aligned}&\mathrm{E}\left(\left(\hat{n}_{f|1}^{(r)}\right)^2\right) \qquad (33)\\ &= \sum_{i=0}^{n-n_f} \mathrm{E}\left(\left(\hat{n}_{f|1}^{(r)}\right)^2 \middle| n_{f|1}^{(r)} = j\right) \cdot \Pr\{n_{f|1}^{(r)} = j\}\\ &\approx \sum_{i=0}^{n-n_f} \left(\frac{\gamma_m^2}{m}+1\right)(j)^2 \cdot \binom{n-n_f}{j}\left(\frac{1}{2w}\right)^j\left(1-\frac{1}{2w}\right)^{n-n_f-j}\\ &= \left(\frac{\gamma_m^2}{m}+1\right)\left(\mathrm{E}((n_{f|1}^{(r)})^2)\right)\\ &= \left(\frac{\gamma_m^2}{m}+1\right)\left((\mathrm{E}(n_{f|1}^{(r)}))^2 + \mathrm{Var}(n_{f|1}^{(r)})\right)\\ &= \left(\frac{\gamma_m^2}{m}+1\right)\left(\left(\frac{1}{2w}(n-n_f)\right)^2 + \frac{1}{2w}\left(1-\frac{1}{2w}\right)(n-n_f)\right) \qquad (34)\end{aligned}$$

To simplify (34), we define one symbol C:

$$C = \mathrm{E}(n_{f|1}^{(r)}) = \frac{1}{2w}(n-n_f), \tag{35}$$

where:

- C is the expected number of flow elements mapped to the alternative estimator.

Applying the two symbols B and C to (12) and (34), we have:

$$\left(\mathrm{E}\left(\hat{n}_{f|1}^{(r)}\right)\right)^2 = C^2, \tag{36}$$

$$\mathrm{E}\left(\hat{n}_{f|1}^{(r)} \cdot \hat{n}_{f|1}^{(r)}\right) = \left(\frac{\gamma_m^2}{m}+1\right)(C^2+B). \tag{37}$$

Then, we can substitute these into (17):

$$\begin{aligned}\mathrm{Var}\left(\hat{n}_{f|1}^{(r)}\right) &= \mathrm{E}\left(\hat{n}_{f|1}^{(r)} \cdot \hat{n}_{f|1}^{(r)}\right) - \left(\mathrm{E}\left(\hat{n}_{f|1}^{(r)}\right)\right)^2 &(38)\\ &= \left(\frac{\gamma_m^2}{m}+1\right)(C^2+B) - C^2 &(39)\\ &= \left(\frac{\gamma_m^2}{m}\right)C^2 + \left(\frac{\gamma_m^2}{m}+1\right)B. &(40)\end{aligned}$$

Finally, the total variance of the per-row estimator (15) is:

$$\begin{aligned}&\mathrm{Var}\left(\hat{n}_f^{(r)}\right)\\ &= \mathrm{Var}\left(\hat{n}_{f|0}^{(r)}\right) + \mathrm{Var}\left(\hat{n}_{f|1}^{(r)}\right)\\ &= \left(\frac{\gamma_m^2}{m}\right)(A^2+C^2) + 2\left(\frac{\gamma_m^2}{m}+1\right)B. \qquad (41)\end{aligned}$$

Estimation Distribution after Aggregation. MOPS generates the final flow spread estimate $\hat{n}_f$ by aggregating the d independent row-wise estimates via the sample median:

$$\hat{n}_f = \mathrm{median}\left(\{\hat{n}_f^{(r)} \mid 0 \le r < d\}\right), \tag{42}$$

where $\hat{n}_f^{(r)}$ denotes the row-wise spread estimator for flow f in the r-th row of the per-flow sketch. We now analyze the statistical properties of $\hat{n}_f$, leveraging the following classical result.

Theorem 2 (CLT for Sample Median)**.** *Let $X_1, X_2, \ldots, X_d$ be i.i.d. real-valued random variables with probability density function $g(x)$, and let s_0 denote their population median, i.e., $G(s_0) = 1/2$ where $G(x)$ is the cumulative distribution function. Then, as $d \to \infty$, the sample median $\hat{s} = \mathrm{median}\{X_1, \ldots, X_d\}$ converges in distribution to:*

$$\hat{s} \sim \mathcal{N}\left(s_0,\ \frac{1}{4d \cdot g(s_0)^2}\right). \tag{43}$$

Previously, we established that the row-wise estimators $\hat{n}_f^{(r)}$ for $0 \le r < d$ are i.i.d. random variables, approximately Gaussian, with expectation given by (13) and variance by (41). Applying Theorem 2, we conclude that the final estimate $\hat{n}_f$ is also asymptotically unbiased:

$$\mathrm{E}(\hat{n}_f) \approx n_f. \tag{44}$$

Moreover, the variance of the sample median can be approximated by:

$$\begin{aligned}\mathrm{Var}(\hat{n}_f) &\approx \frac{1}{4d \cdot g\left(\mathrm{E}[\hat{n}_f^{(r)}]\right)^2}\\ &= \frac{2\pi}{4d} \cdot \mathrm{Var}\left(\hat{n}_f^{(r)}\right)\\ &= \frac{\pi}{2d}\left(\frac{\gamma_m^2}{m}(A^2+C^2) + 2\left(\frac{\gamma_m^2}{m}+1\right)B\right), \qquad (45)\end{aligned}$$

where $\gamma_m \approx 1.04$ for $m \ge 128$ registers in each virtual HLL estimator, and A, B, and C are defined as in above.

From (45), we observe that the standard deviation of $\hat{n}_f$ can be upper bounded as:

$$\mathrm{StdDev}(\hat{n}_f) \le \left(n_f + \frac{1}{2w}n\right) \cdot \sqrt{\frac{\pi}{2d}} \cdot \frac{\gamma_m}{\sqrt{m}}, \tag{46}$$

which satisfies the relative error bound specified in Eq. (4), with $p = \frac{1}{2w}$.