Draft of Theta-Sketch Paper

three authors

ABSTRACT

no abstract yet

1. SINGLE SKETCH

DEFINITION 1.1. *For a stream $\mathcal{S}$ define the "de-duped" stream $D(\mathcal{S})$ as follows: if $\mathcal{S} = s_1 s_2 \ldots s_n$, then for each i, remove all occurrences of s_i from the indices $[i+1, n]$. The resulting stream, where every element appears exactly once, and in the same order as their first appearance in $\mathcal{S}$ is defined to be $D(\mathcal{S})$.*

The following Lemma shows that our distinct count estimator Z for a single stream has the same guarantees as the Morris counter.

LEMMA 1.2. *Let $Z = Z(\mathcal{S}, h)$ be the random variable corresponding to the level when Algorithm 1 is run over stream $\mathcal{S}$ with hash function h. Let $M = M(D(\mathcal{S}), \frac{1}{\alpha}, h)$ be the value of the Morris counter when run over $D(\mathcal{S})$ with base $\frac{1}{\alpha}$ and hash function h. Then for any stream $\mathcal{S}$, the random variables M and Z have identical distribution over the random choice of the hash function.*

[[Add in proof– easy but need to be careful about corner cases.]]

Using the above lemma, and Theorem XXX from Flajolet [], we have the following corollary about the expectation and variance of the estimator Z.

COROLLARY 1.3. *For a stream $\mathcal{S}$, the estimate Z has expectation $E[Z] = u$ and variance $\sigma^2(Z)$ bounded by $\sigma^2(Z) < \frac{n^2}{2k}$. Hence, by choosing $\alpha = \frac{k}{k+1}$ where $k = \frac{1}{2\epsilon^2\delta}$, we have that with probability $1-\delta$,*

$$(1-\epsilon)F_0(S) \leq Z \leq (1+\epsilon)F_0(S).$$

[[We need to add in the bound for the space usage– that is novel.]]

2. NEW ANALYSIS FOR SET OPERATIONS

Suppose we have m streams A_i and B and a set expression $f(\{A_i\})$. We give bounds on the estimator Y for evaluating the expression. Let θ_i be the threshold values obtained for the i^{th} streams using only the first pass and using two independent private hash functions h_i. θ_i, $i \in [1, m]$ are random variables that are independent of each other. We then set $\theta_r = \min \theta_i$.

LEMMA 2.1. *Let $n_{\max} = \max_i |A_i|$ and let $n_f = |f(\{A_i\})|$ The estimate Y is unbiased, i.e. $E[Y] = n_f$. If $\alpha = \frac{k}{k+1}$ where $k \geq \frac{1}{2\epsilon^2\delta}$, then with probability $1-(m+1)\delta$, we have*

$$n_f - \Delta \leq Y \leq n_f + \Delta,$$

where $\Delta = \max(\sqrt{\frac{2(1+\epsilon)n_{\max}}{k} n_f \log(\frac{1}{\delta})}, 1.5(1+\epsilon)\frac{n_{\max}}{k}\log(\frac{1}{\delta}))$.

PROOF. For the given choice of α, for a specific i, with probability $1-\delta$,

$$(1-\epsilon)n_i \leq \frac{k}{\theta_i} \leq (1+\epsilon)n_i,$$

and hence, by taking union bound over all i, with probability $1-m\delta$.

$$\frac{k}{(1+\epsilon)n_i} \leq \theta_i \leq \frac{k}{(1-\epsilon)n_i}.$$

Thus, with probability $1-m\delta$, $\theta = \min_i \theta_i \geq \frac{k}{(1+\epsilon)n_{\max}}$. We now condition on this event happening.

For each element x of $f(A_i)$, let y_x be the indicator variable that indicates whether or not $h(x) < \theta$, i.e. whether or not x is present in the composed sketch. So

$$y_x = \begin{cases} 1 & \text{w.p. } \theta \\ 0 & \text{else.} \end{cases}$$

Let $y = \sum_{i \in f(\{A_i\})} y_i$. Hence $Y = \frac{y}{\theta}$. We then bound y using Chernoff bound.

$$\Pr[|y - n_f\theta| > t] < \exp\left(-\frac{t^2/2}{n_f\theta(1-\theta) + t/3}\right).$$

By choosing $t = \max(\sqrt{n_f\theta(1-\theta)\log(\frac{1}{\delta})}, 1.5\log(\frac{1}{\delta}))$, we have that with probability $1-\delta$,

$$\begin{aligned}|Y - n_f| &< \frac{1}{\theta}\max(\sqrt{n_f\theta(1-\theta)\log(\frac{1}{\delta})}, 1.5\log(\frac{1}{\delta})) \\ &< \max(\sqrt{\frac{2(1+\epsilon)n_{\max}}{k} n_f \log(\frac{1}{\delta})}, 1.5(1+\epsilon)\frac{n_{\max}}{k}\log(\frac{1}{\delta}))\end{aligned}$$

By taking the union bound, the total probability of failure is bounded by $1-(m+1)\delta$. Hence we have the statement of the Lemma.

□

3. SCRATCH SPACE

We then bound the moment generating function $E[\exp(tM)] = \prod_{i \in A\Delta B} E[\exp(tM_i)]$. Also,

$$E[\exp(tM_i)] = \sum_j E[\exp(tM_i)|\theta_{ab} = \alpha_j] P(j|u_a, u_b)$$

where $P(j|u_a, u_b)$ denotes the probability that $\theta_{ab} = \alpha^j$. Also,

$$E[\exp(tM_i) \mid \alpha^j] = 1 + \alpha^j(e^t - 1).$$

The following Lemma follows directly from adapting the Proposition 2 in [] to the case of arbitrary bases.

LEMMA 3.1. *For all i and u,*

$$p_{iu} < \exp(-k)i(1 - \alpha^i)^u.$$

Hence, for $i < \ln(n) - \ln\ln n$,

Similarly, the following Lemma follows from Proposition 4 in [].

LEMMA 3.2. *For $i = 2\log(n) + \delta$, with $\delta \geq 0$, we have $p_{iu} = O(2^{-\delta} n^{-0.99})$.*

Finally, we show the claim. Let bad denote the event that the estimate is $\epsilon|A\delta B|$ away from the the truth. WLOG, let $u_a \geq u_b$. $J_1 = [1, \log(u_a) - \log\log(u_a)]$,

$$\Pr[\mathsf{bad}] \leq \sum_\theta$$