Simpler and Better Cardinality Estimators for HyperLogLog and PCSA*

Seth Pettie
University of Michigan
pettie@umich.edu

Dingyu Wang
University of Michigan
wangdy@umich.edu

Abstract

Cardinality Estimation (aka *Distinct Elements*) is a classic problem in sketching with many industrial applications. Although sketching *algorithms* are fairly simple, analyzing the cardinality *estimators* is notoriously difficult, and even today the state-of-the-art sketches such as HyperLogLog and (compressed) PCSA are not covered in graduate level Big Data courses.

In this paper we define a class of *generalized remaining area* (τ-GRA) estimators, and observe that HyperLogLog, LogLog, and some estimators for PCSA are merely instantiations of τ-GRA for various integral values of τ. We then analyze the limiting relative variance of τ-GRA estimators. It turns out that the standard estimators for HyperLogLog and PCSA can be improved by choosing a *fractional* value of τ. The resulting estimators come *very* close to the Cramér-Rao lower bounds for HyperLogLog and PCSA derived from their Fisher information. Although the Cramér-Rao lower bound *can* be achieved with the Maximum Likelihood Estimator (MLE), the MLE is cumbersome to compute and dynamically update. In contrast, τ-GRA estimators are trivial to update in constant time.

Our presentation assumes only basic calculus and probability, not any complex analysis [FM85, DF03, FFGM07].

1 Introduction

The Problem. A stream $\mathbf{x} = (x_1, \ldots, x_n)$ of elements from a universe $[U]$ is received one at a time. We wish to maintain a small *sketch* S, whose size is independent of n, so that we can return an estimate $\hat{\lambda}$ to the *cardinality* $\lambda = |\{x_1, \ldots, x_n\}|$. Because $\mathbf{x}$ may be partitioned among many machines and processed separately, it is desirable that the resulting sketches be *mergeable*. For this reason we only consider sketches whose state S depends only on the *set* $\{x_1, \ldots, x_n\}$, i.e., it is insensitive to duplicates and is not a function of the *order* in which elements are processed. See [PW21] for a longer discussion of mergeability and [Coh15, Tin14, PWY21] for *non*-mergeable cardinality sketching.

The Model. The Cardinality Estimation/Distinct Elements problem is studied under two models, each with its own conventions. In the RANDOMORACLE model it is assumed that we have access to a uniformly random hash function $h : [U] \to [0, 1]$. By mapping $\mathbf{x}$ to $h(\mathbf{x}) = (h(x_1), \ldots, h(x_n))$,

---

the state of the sketch S can be updated according to a deterministic transition function. In particular, the distribution of the state of S depends only on the cardinality λ, not $\mathbf{x}$. By convention, estimators for sketches in the RANDOMORACLE model are unbiased (or close to unbiased), and their efficiency is measured by the *relative* variance $\lambda^{-2}\mathbb{V}(\hat{\lambda})$, or equivalently, the standard error $\lambda^{-1}\sqrt{\mathbb{V}(\hat{\lambda})}$.[1] The leading constants in the space usage and variance are typically stated explicitly. See [FM85, Fla90, DF03, FFGM07, Gir09, CG06, EVF06, BGH+09, Lan17, Lum10, PW21, ŁU22, Oha21].

In the STANDARD model we can generate independent random bits, but must explicitly store any hash functions. By convention, the estimators in this model come with an (ϵ, δ)-guarantee (rather than bias and variance guarantees), i.e., $\mathbb{P}(\hat{\lambda} \notin [(1-\epsilon)\lambda, (1+\epsilon)\lambda]) \le \delta$. The space depends on ϵ, δ, U, and is expressed in big-Oh notation, often with large hidden constants. In this model $\Theta(\epsilon^{-2}\log\delta^{-1}+\log U)$ bits of space is necessary and sufficient. See Jayram and Woodruff [JW13] and Alon, Matias, and Szegedy [AMS99] for the lower bound and Błasiok [Bł20] for the upper bound. See also [KNW10, GT01, BJK+02, BKS02, IW03] for other results in the STANDARD model.

In this paper we assume the RANDOMORACLE model. The sketches used in practice (HyperLogLog, PCSA, k-Min, etc.) all originate in the RANDOMORACLE model and despite being implemented with imperfect hash functions, their empirical behavior closely matches their theoretical analysis [The19, HNH13, Lan17].

Sketches and Estimators. In 1983 Flajolet and Martin [FM85] developed the first non-trivial sketch called Probabilistic Counting with Stochastic Averaging (PCSA).[2] A PCSA sketch S_{PCSA} consists of an array of m bit vectors or *subsketches*. The random oracle produces a pair $(h, g)(x)$, where $h(x) \in [m]$ is a uniformly random subsketch index and $g(x) \in \mathbb{Z}^+$ is equal to k with probability 2^{-k}. The bit $S_{\mathrm{PCSA}}(j, k)$ is 1 if there exists an x_i in the stream with $h(x_i) = j$ and $g(x_i) = k$, and 0 otherwise. Define $z(j) = \min\{k : S_{\mathrm{PCSA}}(j, k) = 0\}$ to be the position of the least significant zero in the jth subsketch. Each $z(j)$ is individually a decent estimate of $\log(\lambda/m)$. Flajolet and Martin [FM85] analyzed the "first zero" estimator for PCSA, namely

$$\hat{\lambda}_{\mathrm{FM}}(S_{\mathrm{PCSA}}) \propto m \cdot 2^{\frac{1}{m}\sum_{j=1}^{m} z(j)}$$

and proved it has relative variance about $0.6/m$ and hence standard error about $0.78/\sqrt{m}$. It suffices to keep $\log U$ bits per subsketch, so PCSA requires $m\log U$ bits. Although the "first zero" has better concentration than the "last one," the latter is much cheaper to store. In 2003 Durand and Flajolet [DF03] implemented this idea in the LogLog sketch S_{LL}, which requires only $m\log\log U$ bits.

$$S_{\mathrm{LL}}(j) = \max\{k : S_{\mathrm{PCSA}}(j, k) = 1\}.$$

Durand and Flajolet proved that the estimator

$$\hat{\lambda}_{\mathrm{DF}}(S_{\mathrm{LL}}) \propto m \cdot 2^{\frac{1}{m}\sum_{j=1}^{m} S_{\mathrm{LL}}(j)}$$

---

[1] We use $\mathbb{P}, \mathbb{E}$, and $\mathbb{V}$ for probability mass, expectation, and variance.

[2] Although this paper is seminal — it kicked off the field of statistical analysis of data streams — it seems to be widely unread. It was cited in a paper by Estan, Varghase, and Fisk [EVF06], who *reinvented* PCSA under the name Multiresolution Bitmap. It is often cited (falsely) as the paper introducing the (Hyper)LogLog sketch and the k-Min sketch.

has relative variance about C_{DF}/m and standard error about $\sqrt{C_{\mathrm{DF}}/m} \approx 1.3/\sqrt{m}$, where $C_{\mathrm{DF}} = \frac{2\pi^2 + \log^2 2}{12} < 1.69$.[3] This estimator can be regarded as taking the *geometric mean* of individual estimates $2^{S_{\mathrm{DF}}(1)}, \ldots, 2^{S_{\mathrm{DF}}(m)}$. In 2007, Flajolet, Fusy, Gandouet, and Meunier [FFGM07] proposed a better estimator for LogLog based on the *harmonic* mean:

$$\hat{\lambda}_{\mathrm{FFGM}}(S_{\mathrm{LL}}) \propto m^2 \cdot \left(\sum_{j=1}^{m} 2^{-S_{\mathrm{LL}}(j)} \right)^{-1}$$

and called the resulting sketch HyperLogLog. It has relative variance roughly C_{FFGM}/m and standard error $\sqrt{C_{\mathrm{FFGM}}/m} \approx 1.04/\sqrt{m}$, where $C_{\mathrm{FFGM}} = 3\ln 2 - 1 \approx 1.07944$. (The constants C_{DF} and C_{FFGM} are, in fact, limiting constants as $m \to \infty$.)

Optimal Cardinality Sketching. The sketches above consist of m subsketches, where the memory scales linearly with m, and the relative variance with m^{-1}. The most reasonable way to measure the *overall efficiency* of a sketch is by its memory-variance product (MVP). Scheuermann and Mauve [SM07] experimented with *compressed* versions of PCSA and (Hyper)LogLog,[4] and found Compressed-PCSA to be slightly MVP-superior to Compressed-HyperLogLog. Lang [Lan17] also experimented with these compressed sketches, but used *maximum likelihood* estimators (MLE) instead.[5] He found that using MLE, Compressed-PCSA is *substantially* better than Compressed-HyperLogLog. In general, the MLE $\hat{\lambda}_{\mathrm{MLE}}(S)$ of a sketch S is the λ^* that maximizes the probability of seeing S, conditioned on $\lambda = \lambda^*$ being the true cardinality. The MLE is cumbersome to compute and update. Lang [Lan17] also found that a simple "coupon collector" estimator based on counting the number of 1s in a PCSA sketch gives better estimates than Flajolet and Martin's original estimator $\hat{\lambda}_{\mathrm{FM}}$.

$$\hat{\lambda}_{\mathrm{Lang}}(S_{\mathrm{PCSA}}) \propto m \cdot 2^{\frac{1}{m} \sum_{j=1}^{m} \sum_{k \geq 1} S_{\mathrm{PCSA}}(j,k)}.$$

Lang [Lan17] argued informally that the relative variance of $\hat{\lambda}_{\mathrm{Lang}}$ should be about $(\log^2 2)/m$, which agreed with his experiments.

One annoying feature of all the sketches cited above is that their relative variance (and bias) are not fixed but *multiplicatively periodic* with period factor 2. (If PCSA and LogLog were defined in base q they would be multiplicatively periodic with period q.) The magnitude of these periodic functions is tiny, but *independent* of m. Pettie and Wang [PW21] gave a generic "smoothing" mechanism to get rid of this periodic behavior. They formally studied the optimality of sketches under the memory-variance product (MVP), where both "memory" and "variance" are interpreted as taking on their information-theorically optimum values. They defined the *Fish-number* of a sketch in terms of (1) its <u>Fi</u>sher information, which controls the variance of an optimal estimator (e.g., MLE is asymptotically optimal), and (2) its <u>Sh</u>annon entropy, which controls its memory under optimal compression. They found closed form expressions for the entropy and Fisher information of base-q variants of PCSA and LogLog, and discovered that q-PCSA has Fish-number $H_0/I_0 \approx 1.98$ for all q, and q-LogLog has a Fish-number strictly larger than H_0/I_0, but that it tends to H_0/I_0 in the

---

[3]All logarithms are natural unless specified otherwise.

[4]It is straightforward to show that the entropy of both sketches is $O(m)$ bits.

[5]Lang [Lan17] formulated this as Minimum Description Length (MDL) estimation, which is equivalent in this context.

limit, as $q \to \infty$. Here H_0 and I_0 are precisely defined constants.[6] The Fishmonger sketch of [PW21] is a smoothed, entropy compressed version of PCSA with an MLE estimator, which achieves $1/\sqrt{m}$ standard error with $(1 + o(1))mH_0/I_0$ bits of space. Moreover, they give compelling evidence that Fishmonger is optimal, i.e., no sketch can achieve Fish-number (memory-variance product) better than H_0/I_0.[7] For example, to achieve 1% standard error, [PW21] indicates that one needs $(H_0/I_0)(0.01)^2$ bits, which is about 2.42 kilobytes.

1.1 Dartboards and Remaining Area

Ting [Tin14] introduced a very intuitive *visual* way to think about cardinality sketches he called the *area cutting process*. Pettie, Wang, and Yin [PW21, PWY21] described a constrained version of Ting's process they called the *Dartboard* model.[8] The elements of this model are as follows:

Dartboard and Darts. The *dartboard* is a unit square $[0, 1]^2$. When an element (dart) $x \in [U]$ arrives, it is *thrown* at a point $h(x) \in [0, 1]^2$ in the dartboard determined by the random oracle h.

Cells and States. The dartboard is partitioned into a countable set $\mathcal{C}$ of *cells*. Every cell may be *occupied* or *free*. The *state* of the sketch is defined by the set $\sigma \subseteq \mathcal{C}$ of occupied cells. The *state space* is some subset of $2^{\mathcal{C}}$.

Occupation Rules. If a dart is thrown at an occupied cell, the state does not change. If a dart is thrown at a free cell c, and the current state is σ, the new state is $f(\sigma, c) \supseteq \sigma \cup \{c\}$ in the state space.

Observe that the state transition function $f(\sigma, c)$ may force a cell to become *occupied* even though it contains no dart, which occurs in (Hyper)LogLog, for example. See Figure 1.
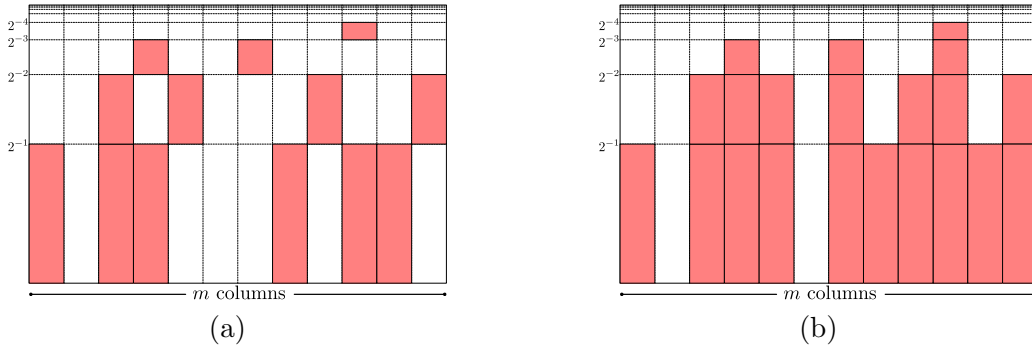

Figure 1: The cell partition used by PCSA and (Hyper)LogLog. (a) A possible state of PCSA. Occupied (red) cells are precisely those containing darts. (b) The corresponding state of (Hyper)LogLog. Occupied (red) cells contain a dart, or lie below a cell in the same column that contains a dart.

---

[6] $I_0 = \pi^2/6$ measures the Fisher information and $H_0 = \frac{1}{\log 2} + \sum_{k=1}^{\infty} \frac{1}{k} \log_2 (1 + 1/k)$ the Shannon entropy of a PCSA sketch.

[7] The optimum Fish-number in the class of "linearizable" sketches is H_0/I_0, and all the popular sketches are linearizable, such as HyperLogLog, PCSA, k-Min, etc. Known sketches that fail to be linearizable are for subtle technical reasons, e.g., AdaptiveSampling [Fla90, GT01].

[8] The two are essentially identical, except that Ting's model does not have an explicit state space, and allows for non-deterministic state transitions.

4

It was observed [Tin14, PW21] that the Dartboard model includes all mergeable sketches, and even some non-mergeable ones like the S-Bitmap [CCSN11].[9] A useful summary statistic of state σ is its *remaining area*

$$\text{RemainingArea}(\sigma) = \sum_{c \in \mathcal{C} \setminus \sigma} |c|,$$

where $|c|$ is the size of cell c. In other words, the remaining area is the total size of all free cells, or equivalently, the probability that the sketch changes upon seeing the next *distinct* element. Remaining area plays a key role in the (non-mergeable) Martingale sketches of [Coh15, Tin14, PWY21]. It also gives us a less fancy way to describe the HyperLogLog estimator without mentioning harmonic means:

$$\hat{\lambda}_{\text{FFGM}}(S_{\text{LL}}) \propto m \left(\text{RemainingArea}(S_{\text{LL}})\right)^{-1}.$$

Estimating the cardinality proportional to the reciprocal of the remaining area is reasonable for *any* sketch. This is the optimal estimator for k-Min-type sketches [CG06, Lum10], and as we will see, superior to Flajolet and Martin's original $\hat{\lambda}_{\text{FM}}$ estimator for PCSA.

Generalized Remaining Area. Rather than have each cell $c \notin \sigma$ contribute $|c|$ to the remaining area of state σ, we could let it contribute $|c|^{\tau}$ instead for some fixed exponent $\tau > 0$. The resulting summary statistic is called τ-*generalized remaining area*, or τ-GRA.

$$\tau\text{-GRA}(\sigma) = \sum_{c \in \mathcal{C} \setminus \sigma} |c|^{\tau}.$$

Note that 0-GRA counts the number of free cells, which we regard as equivalent to counting the number of occupied cells, as is done explicitly by $\hat{\lambda}_{\text{Lang}}$.[10] It is also possible to analyze τ-GRA when $\tau < 0$ by summing over occupied cells rather than free cells. However, in this paper we focus only on $\tau \geq 0$.

1.2 New Results

A conceptual contribution of this paper is the introduction of the τ-GRA summary statistic. The main technical contribution is a relatively simple analysis of the limiting efficiency of estimators for PCSA and (Hyper)LogLog based on the τ-GRA statistic. Our analysis has several benefits.

A Unified View. We show that HyperLogLog is based on 1-GRA and that, properly interpreted, LogLog is based on 0-GRA. Moreover, Lang's "coupon collector" estimator $\hat{\lambda}_{\text{Lang}}$ for PCSA is based on 0-GRA. Our analysis confirms Lang's back-of-the-envelope calculations that $\hat{\lambda}_{\text{Lang}}$ has limiting relative variance $(\log^2 2)/m$.

Simplicity. We use two techniques to dramatically simplify the analysis of τ-GRA-based estimators. The first, which has been used before [FM85, FFGM07, PW21, PWY21], is to consider a "Poissonized" dartboard model, which allows us to avoid issues with small cardinalities and

---

[9](In principle the dartboard can be partitioned into cells consisting of individual hash values. Occupied cells are *by definition* hash values that have no effect on the state. The requirement that the sketch be insensitive to duplicates demands that all cells hit by darts become occupied.)

[10]In our stylized model there are an infinite number of cells, but in practice there are a finite number, so counting free cells is equivalent to counting occupied cells.

infinitesimal negative correlations between cells. The second is a smoothing operation similar to the one introduced in [PW21]. The combined effect of Poissonization and smoothing is to make the sketch truly scale-invariant at every cardinality, without any periodic behavior.

Efficiency. A *statistically* optimal estimator for PCSA or LogLog meets the Cramér-Rao lower bound, which depends on the Fisher information of the given sketch; see [PW21]. It is known [CB02, Vaa98] that the maximum likelihood estimator $\hat{\lambda}_{\text{MLE}}$ meets the Cramér-Rao lower bound asymptotically, as $m \to \infty$, but MLE is not particularly simple to update as the sketch changes. The limiting relative variance of HyperLogLog's $\hat{\lambda}_{\text{FFGM}}$ is $(3\log 2 - 1)/m \approx 1.07944/m$, plus a tiny periodic function. Pettie and Wang's analysis [PW21, Lemmas 4,5] shows that the the Cramér-Rao lower bound for (Hyper)LogLog is $\frac{\log 2}{\pi^2/6-1}/m \approx 1.07475/m$, which does not leave much room for improvement! In contrast, there is a wider gap between the limiting variance of PCSA's $\hat{\lambda}_{\text{DF}}$, namely $0.6/m$, or Lang's improvement $\hat{\lambda}_{\text{Lang}}$, namely $(\log^2 2)/m \approx 0.48/m$, and the Cramér-Rao lower bound [PW21, Theorem 3] of $\frac{\pi^2}{6\log 2}/m \approx 0.42138/m$. By choosing the optimal τs, our τ-GRA-based estimators achieve relative variance $1.0750/m$ for the LogLog sketch and $0.435532/m$ for the PCSA sketch, in both cases *nearly closing the gap* between the best known explicit estimators and the Cramér-Rao lower bound. The improvement to HyperLogLog is probably not worth implementing, but the improvement to PCSA can lead to an immediate improvement to practical implementations of PCSA, e.g., the CPC (Compressed Probabilistic Counting) sketch included in Apache *DataSketches* [The19].

Figures 2 and 3 illustrate the efficiency of τ-GRA-based estimators relative to other estimators, and Table 1 summarizes the same information symbolically.
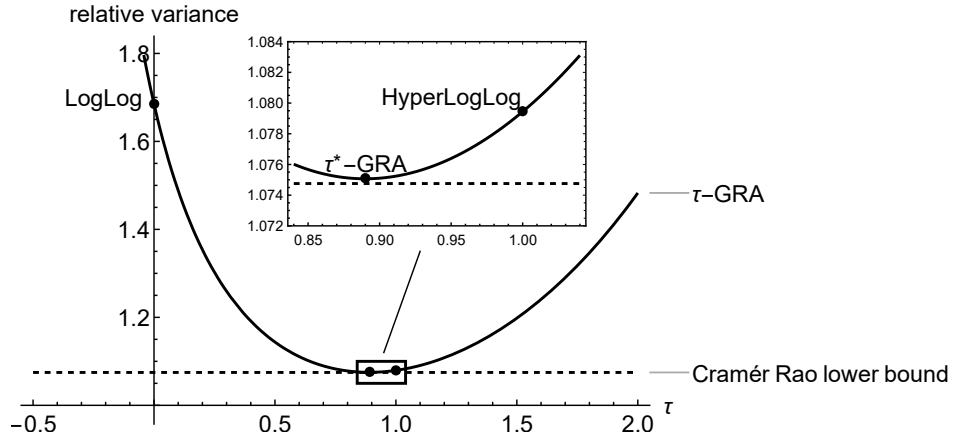

Figure 2: Relative variance of estimators for the LogLog sketch. The τ-GRA estimator attains minimum variance at $\tau^* = 0.88989$, which comes within 0.02% of the Cramér-Rao lower bound. As a comparison, HyperLogLog is 0.4% over the bound.

1.3 Related Work

One weakness of HyperLogLog is its poor performance on small cardinalities $\lambda = \tilde{O}(m)$. Heule et al. [HNH13] proposed improvements to [FFGM07]'s estimator on small cardinalities, as well as
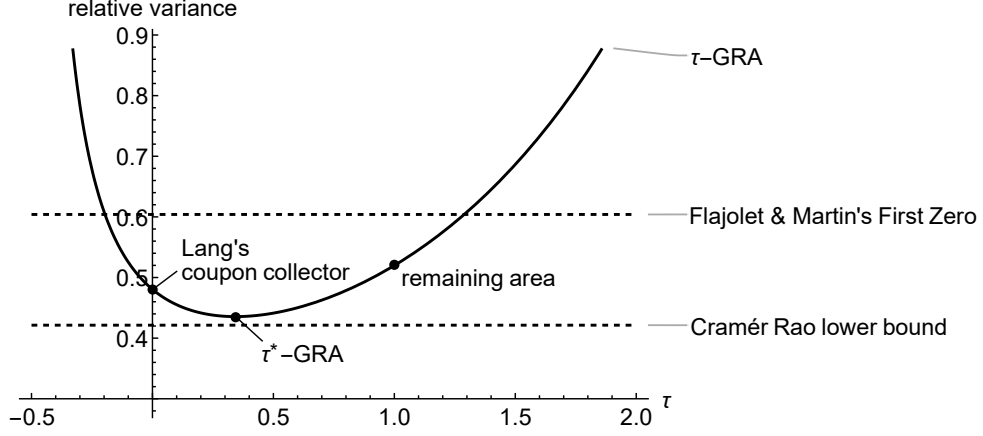
Figure 3: Relative variance of estimators for the PCSA sketch. The τ-GRA estimator attains minimum variance at $\tau^* = 0.343557$, which comes within 3% of the Cramér-Rao lower bound.

SKETCH & ESTIMATOR		LIMITING RELATIVE VARIANCE	CITATION
PCSA Flajolet & Martin 1983			[FM85]
First Zero ($\hat{\lambda}_{\mathrm{FM}}$)		$\approx 0.6/m + \theta(\lambda)$	[FM85]
Coupon Collector ($\hat{\lambda}_{\mathrm{Lang}}$)		$\approx (\log^2 2)/m + \theta(\lambda) \approx 0.48/m$	[Lan17]
Smoothed 0-GRA		$(\log^2 2)/m$	**Theorem 4**
Smoothed 1-GRA		$\frac{3\ln 2}{4}/m \approx 0.51986/m$	**Theorem 4**
Smoothed τ-GRA $\quad \tau = 0.343557$		$\approx 0.435532/m$	**Theorem 4**
MLE / Cramér-Rao Lower Bound		$\frac{\pi^2}{6\log 2}/m \approx 0.42138/m$	[PW21]

LogLog Durand and Flajolet 2003			[DF03]
Geometric Mean ($\hat{\lambda}_{\mathrm{DF}}$)		$\frac{2\pi^2+\log^2 2}{12}/m + \theta(\lambda) \approx 1.69/m$	[DF03]
Harmonic Mean ($\hat{\lambda}_{\mathrm{FFGM}}$)		$(3\log 2 - 1)/m + \theta(\lambda) \approx 1.07944/m$	[FFGM07]
Smoothed 0-GRA		$\frac{2\pi^2+\log^2 2}{12}/m \approx 1.69/m$	**Theorem 3**
Smoothed 1-GRA		$(3\log 2 - 1)/m \approx 1.07944/m$	**Theorem 3**
Smoothed τ-GRA $\quad \tau = 0.889897$		$\approx 1.07507/m$	**Theorem 3**
MLE / Cramér-Rao Lower Bound		$\frac{\log 2}{\pi^2/6-1}/m \approx 1.07475/m$	[PW21]

Table 1: Relative variance as $m, \lambda \to \infty$. All $\theta(\lambda)$ functions are multiplicatively periodic with period 2, which have a small magnitude independent of m. The "smoothing" mechanism (Section 2) eliminates periodic behavior.

some more efficient sketch encodings when λ is small. Ertl [Ert17] experimented with maximum likelihood estimation (MLE) for HyperLogLog sketches, which behaves well at all cardinalities.

Łukasiewicz and Uznański [ŁU22] developed a HyperLogLog-like sketch that, in our terminology, samples $g(x)$ from a *Gumbel* distribution rather than a *geometric* distribution. As the maximum of several Gumbel-distributed variables is Gumbel-distributed, this resulted in a simpler analysis relative to [FFGM07].

It is well known that the entropy of HyperLogLog is $O(m)$. Durand [Dur04] gave a prefix-

7

free code for (Hyper)LogLog with expected length $3.01m$, and Pettie and Wang [PW21] gave a precise expression for the entropy of (Hyper)LogLog, which is about $2.83m$. Xiao et al. [XCZL20] proposed lossy compressions of HyperLogLog to $4m$ and even $3m$ bits, but their variance calculation is incorrect; see [PW21] for a discussion of the problems of lossy compression in this context. Very recently Karppa and Pagh [KP22] presented a lossless compression of HyperLogLog to $(1 + o(1))m \log \log \log U$ bits ('HyperLogLogLog') while still allowing fast update times.

Pettie, Wang, and Yin [PW21] proposed a class of *Curtain* sketches that combine elements of LogLog and PCSA while being easily compressible, but they only analyzed them in the *non-mergeable* setting of [Coh15, Tin14]. Ohayon [Oha21] analyzed the most practical (and mergeable) Curtain$(2, \infty, 1)$ sketch, and found it to be substantially more efficient than HyperLogLog in terms of memory-variance product. In particular, its limiting variance is C/m, $C = \frac{41 \log 2}{16} - 1 \approx 0.776$ while using only m more bits than HyperLogLog or any lossless compression thereof, e.g. [The19] or [KP22].

1.4 Organization

In Section 2 we define *scale-invariance*, and introduce *Poissonization* and *smoothing* mechanisms to achieve scale-invariance. Section 3 introduces cardinality estimates based on the τ-GRA statistic, and Sections 4 and 5 analyze the behavior of these estimators on the LogLog and PCSA sketches, respectively. We conclude with some remarks in Section 6.

2 Poissonization and Smoothing

Suppose we have an estimator E_λ at cardinality λ. Ideally, for a statistic to be a *measurement* of cardinality the relative error should distribute identically for any cardinality, i.e., E_λ should be *scale-invariant*.

Definition 1 (scale-invariance)**.** Let E_λ be an estimator of λ. We say E_λ is *scale-invariant* if for any $\lambda > 0$,

$$\frac{E_\lambda}{\lambda} \sim E_1.$$

Note that scale-invariance implies unbiasedness and constant relative variance that is independent of λ. Since $\mathbb{E}E_\lambda = \lambda \mathbb{E}E_1$, if $\mathbb{E}E_1 \neq 1$ then we can replace E_λ with $\frac{E_\lambda}{\mathbb{E}E_1}$ to make it an unbiased estimate of λ. Moreover, $\mathbb{V}(E_\lambda) = \lambda^2 \mathbb{V}(E_1)$, where $\mathbb{V}(E_1)$ is some fixed constant independent of λ.

Much of the simplicity and elegance of our analysis relies on beginning from *this* definition of strict scale-invariance. Unfortunately, in the real world the PCSA and HyperLogLog sketches are only *approximately* scale-invariant, stemming from two causes discussed in the introduction.

- Fixed-size sketches have two "edge effects," when $\lambda = \tilde{O}(m)$ is small and when $\lambda = \tilde{\Omega}(U)$ is approaching the size of the universe. The latter problem cropped up when $U = 2^{32}$ was small [FFGM07] but is generally not an issue when $U \geq 2^{64}$. See [FFGM07, HNH13, Ert17] for improved estimation methods for small cardinalities.

- Sketches that store continuous random variables (like k-Min [Coh97, Gir09, CG06, Lum10]) exhibit no periodic behavior but sketches that discretize their data are multiplicatively periodic in the base of the discretization, which is 2 in the case of standard PCSA and (Hyper)LogLog.

Therefore, the relative variance of sketches like PCSA and HyperLogLog are *not* actually fixed constants independent of λ but periodic functions of λ, whose magnitude is small but *independent* of m, the size of the sketch.

We consider a *smoothed*, *Poissonized*, and *infinite* dartboard model to make the task of variance analysis simpler.

Definition 2 (Smoothed, Poissonized, Infinite model)**.** The dartboard model and cell partition of PCSA and (Hyper)LogLog are changed as follows.

Smoothing. The sketch consists of m subsketches; these correspond to the columns in Figure 1. Pick a vector $\mathbf{R} = (R_1, \ldots, R_m)$ of *offsets*. Cell j in column i now covers the vertical interval $(2^{-j-R_i}, 2^{-(j-1)-R_i}]$. We will pick $\mathbf{R}$ in two ways as is convenient. In Section 4 we choose each $R_i \in [0, 1)$ uniformly at random, independent of other offsets. In Section 5 we choose $\mathbf{R} = (0, 1/m, 2/m, \ldots, (m-1)/m)$ to be the uniformly spaced offset vector.

Infinite Dartboard. Rather than index cells by $\mathbb{Z}^+$, index them by $\mathbb{Z}$, i.e., the dartboard has unit width and infinite height. For example, cell 0 covers the vertical interval $(2^{-R_i}, 2^{1-R_i}]$, cell -5 covers the vertical interval $(2^{5-R_i}, 2^{6-R_i}]$, etc.

Poissonization. In the usual dartboard, the probability that a cell c remains free at cardinality λ is $(1 - |c|)^\lambda \to e^{-|c|\lambda}$ and the correlation between cells vanishes as $\lambda \to \infty$. For simplicity, these asymptotic properties can be achieved even for small λ with *Poissonization*. Informally speaking, with Poissonization, for each insertion, instead of throwing *one* dart at the board, darts *appear* on the board memorylessly with density 1. Formally speaking, for every new insertion, a Poisson point process on the infinite board with density 1 is added to the board, where each point in the process corresponds to a dart. Thus, after λ insertions, the darts on the board form a Poisson point process with density λ. By construction, for any λ—even $\lambda = 1$—the cells are independent and a cell c will remain free with probability precisely $e^{-|c|\lambda}$.
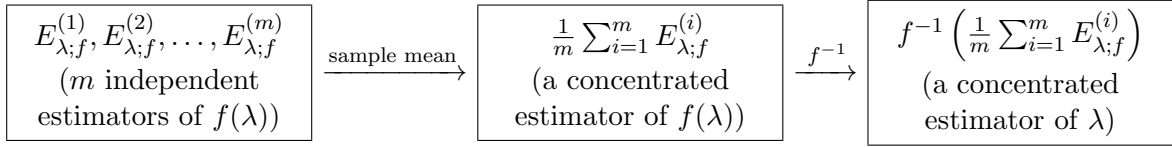
Smoothing eliminates periodic behavior, and the combination of Poissonization and the Infinite Dartboard makes the distribution of the sketch scale invariant for all λ. Our justification for these changes is that they make the analysis simpler, and it does not really matter whether they are implemented in practice once λ is not too small. For example, w.h.p., there is no way to detect whether we are in a unit or infinite dartboard once $\lambda = \Omega(m \log m)$ as all cells indexed by $\mathbb{Z} - \mathbb{Z}^+$ will be occupied. Moreover, as $\lambda \to \infty$ the distribution of the true dartboard converges toward the Poissonized dartboard. Smoothing eliminates the tiny periodic behavior of the estimator, but these effects are too small to worry about unless the magnitude of this periodic function is close to the desired variance, in which case smoothing *should* be implemented in practice.

Remark 1. Pettie and Wang [PW21] introduced smoothing in 2021 to make the analysis of Fishnumbers well defined and non-periodic. Łukasiewicz and Uznański [ŁU22] used smoothing to reduce the space of their cardinality sketch from $O(m(\log \log U + \log \epsilon^{-1}))$ bits to $O(m \log \log U)$ bits, matching HyperLogLog asymptotically.

From this point on, the smoothed, Poissonized, and infinite dartboard model is assumed.

3   Estimation by Generalized Remaining Area

Cardinality estimation can be viewed as a *point estimation* problem where the number of subsketches is the number of independent samples/observations. Classically, one can produce i.i.d. estimates $\left(E_\lambda^{(i)}\right)_{i\in[1,m]}$ of λ with each subsketch and then use the sample mean as the combined estimator. A more general framework is to produce estimates $\left(E_{\lambda;f}^{(i)}\right)_{i\in[1,m]}$ of $f(\lambda)$ for some monotonic function f, then take the sample mean $\frac{1}{m}\sum_{i=1}^m E_{\lambda;f}^{(i)}$, which is concentrated around $f(\lambda)$. Thus we can recover an estimator of λ by applying f^{-1} to the sample mean. This process is summarized as follows.

$$\boxed{\begin{array}{c} E_{\lambda;f}^{(1)}, E_{\lambda;f}^{(2)}, \ldots, E_{\lambda;f}^{(m)} \\ (m \text{ independent} \\ \text{estimators of } f(\lambda)) \end{array}} \xrightarrow{\text{ sample mean }} \boxed{\begin{array}{c} \frac{1}{m}\sum_{i=1}^m E_{\lambda;f}^{(i)} \\ (\text{a concentrated} \\ \text{estimator of } f(\lambda)) \end{array}} \xrightarrow{f^{-1}} \boxed{\begin{array}{c} f^{-1}\left(\frac{1}{m}\sum_{i=1}^m E_{\lambda;f}^{(i)}\right) \\ (\text{a concentrated} \\ \text{estimator of } \lambda) \end{array}}$$

An important example is its application to the *remaining area*. The remaining area (of one subsketch) offers a natural estimate for λ^{-1}. One can get a concentrated estimation for λ^{-1} using the sample mean of remaining areas of the subsketches and then take the inverse to get a concentrated estimation for λ. This is exactly what the HyperLogLog estimator $\hat{\lambda}_{\text{FFGM}}$ does.

The remaining area estimates λ^{-1} and in general, the τ-generalized remaining area estimates $\lambda^{-\tau}$. Let $A_{\lambda;\tau}$ be the τ-generalized remaining area of one subsketch and $A_{\lambda;\tau}^{(1)}, A_{\lambda;\tau}^{(2)}, \ldots, A_{\lambda;\tau}^{(m)}$ be m i.i.d. copies. Thus by the same process, we get a generic estimator $\hat{\lambda}_{\tau;m}$ based on τ-GRA.

$$\hat{\lambda}_{\tau;m} \propto \left(\frac{1}{m}\sum_{i=1}^m A_{\lambda;\tau}^{(i)}\right)^{-\tau^{-1}}.$$

For any sketch, it turns out that the induced estimator $\hat{\lambda}_{\tau;m}$ is scale-invariant if the τ-GRA statistic itself is τ-*scale-invariant*.

Definition 3 (τ-scale-invariance). Let $A_{\lambda;\tau}$ be the τ-generalized remaining area of a sketch. We say $A_{\lambda;\tau}$ is τ-scale-invariant if $A_{\lambda;\tau} \sim \lambda^{-\tau} A_{1;\tau}$ for any $\lambda > 0$.

Theorem 1. *If $A_{\lambda;\tau}$ is τ-scale-invariant, then*

$$\hat{\lambda}_{\tau;m}^* = \left(\frac{1}{m}\sum_{i=1}^m A_{\lambda;\tau}^{(i)}\right)^{-\tau^{-1}}.$$

is a scale-invariant estimator for λ.

Proof. By default, $\hat{\lambda}_{\tau;m}^*$ is the estimator at cardinality λ. When needed, we use $\hat{\lambda}_{\tau;m}^*[\lambda']$ to indicate that it is being evaluated on a sketch with cardinality λ'. By the τ-scale-invariance of $A_{\lambda;\tau}$, we have $A_{\lambda;\tau} \sim \lambda^{-\tau} A_{1;\tau}$. Thus

$$\hat{\lambda}_{\tau;m}^*[\lambda] \sim \left(\frac{1}{m}\sum_{i=1}^m \lambda^{-\tau} A_{1;\tau}^{(i)}\right)^{-\tau^{-1}} = \left(\frac{1}{m}\lambda^{-\tau}\sum_{i=1}^m A_{1;\tau}^{(i)}\right)^{-\tau^{-1}} = \lambda \cdot \hat{\lambda}_{\tau;m}^*[1].$$

$\square$

Since we care about the asymptotic region, we prove the following useful theorem that expresses the asymptotic mean and variance of $\hat{\lambda}^*_{\tau;m}$ by the mean and variance of $A_{1;\tau}$ as $m \to \infty$. Note that although $\hat{\lambda}^*_{\tau;m}$ is scale-invariant, it is not yet normalized to be unbiased; the estimator $\hat{\lambda}_{\tau;m}$ will be the unbiased version of $\hat{\lambda}^*_{\tau;m}$. The asymptotic relative variance after normalization is also given in the lemma.

Theorem 2. *If $A_{1;\tau}$ is τ-scale-invariant with finite variance, we have for any $\lambda > 0$,*

1. $\lim\limits_{m \to \infty} \mathbb{E}\hat{\lambda}^*_{\tau;m} = \lambda(\mathbb{E}A_{1;\tau})^{-\tau^{-1}}.$

2. $\lim\limits_{m \to \infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}^*_{\tau;m}) = \tau^{-2}(\mathbb{E}A_{1;\tau})^{-2\tau^{-1}-2}\mathbb{V}(A_{1;\tau}).$

3. *For any $\lambda > 0$, the normalized estimator $\hat{\lambda}_{\tau;m} = (\mathbb{E}A_{1;\tau})^{\tau^{-1}}\hat{\lambda}^*_{\tau;m}$ is asymptotically unbiased and has limit relative variance*

$$\lim\limits_{m \to \infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \tau^{-2}(\mathbb{E}A_{1;\tau})^{-2}\mathbb{V}(A_{1;\tau}).$$

Proof. By scale-invariance, it suffices to consider the case $\lambda = 1$. Let $X = A_{1;\tau}$ and $Y_m = \frac{1}{m}\sum_{i=1}^{m} A_{1;\tau}^{(i)}$ be the mean of m copies of X. Define $f(x) = x^{-\tau^{-1}}$. Then $\hat{\lambda}^*_{\tau;m} = f(Y_m)$. Since we consider the case as $m \to \infty$, by the central limit theorem, Y_m is asymptotically normal around $\mathbb{E}X$. With high probability we have $Y_m \in (\mathbb{E}X - \frac{\log m}{\sqrt{m}}, \mathbb{E}X + \frac{\log m}{\sqrt{m}})$. Consider the first order approximation in this small neighborhood.

$$f(x) = f(\mathbb{E}X) + (x - \mathbb{E}X)f'(\mathbb{E}X) + O((x - \mathbb{E}X)^2)$$
$$= (\mathbb{E}X)^{-\tau^{-1}} - (x - \mathbb{E}X)\tau^{-1}(\mathbb{E}X)^{-\tau^{-1}-1} + O((x - \mathbb{E}X)^2). \tag{1}$$

Note that $\mathbb{E}Y_m = \mathbb{E}X$ and $\mathbb{V}(Y_m) = \frac{1}{m}\mathbb{V}(X) = O(\frac{1}{m})$. Then we have

$$\mathbb{E}f(Y_m) = (\mathbb{E}X)^{-\tau^{-1}} - (\mathbb{E}Y_m - \mathbb{E}X)\tau^{-1}(\mathbb{E}X)^{-\tau^{-1}-1} + O(\mathbb{V}(Y_m))$$
$$= (\mathbb{E}X)^{-\tau^{-1}} + O(\tfrac{1}{m}). \tag{2}$$

Turning now to the variance,

$$\mathbb{V}(f(Y_m)) = \mathbb{E}\left(f(Y_m) - \mathbb{E}f(Y_m)\right)^2$$
$$= \mathbb{E}\left((Y_m - \mathbb{E}X)\tau^{-1}(\mathbb{E}X)^{-\tau^{-1}-1} + O(\tfrac{1}{m})\right)^2 \qquad \text{(1) and (2)}$$
$$= \mathbb{V}(Y_m)\tau^{-2}(\mathbb{E}X)^{-2\tau^{-1}-2} + O(\tfrac{1}{m^2}),$$

where we note that $\mathbb{E}(Y_m - \mathbb{E}X)^2 = \mathbb{V}(Y_m)$ and $\mathbb{E}(Y_m - \mathbb{E}X) = 0$. As $\mathbb{V}(Y_m) = \frac{1}{m}\mathbb{V}(X)$, this implies that the normalized variance is

$$m\mathbb{V}(f(Y_m)) = \mathbb{V}(X)\tau^{-2}(\mathbb{E}X)^{-2\tau^{-1}-2} + O(\tfrac{1}{m}). \tag{3}$$

We can now obtain a strictly unbiased estimator $(\mathbb{E}\hat{\lambda}^*_{\tau;m}[1])^{-1}\hat{\lambda}^*_{\tau;m}$, where $\hat{\lambda}^*_{\tau;m}[1]$ is the output of the estimator at cardinality (density) 1. We do not know precisely what $\mathbb{E}\hat{\lambda}^*_{\tau;m}[1]$ is, but

$\lim_{m\to\infty} \mathbb{E}\hat{\lambda}^*_{\tau;m}[1] = (\mathbb{E}A_{1;\tau})^{-\tau^{-1}}$, so $\hat{\lambda}_{\tau;m} = (\mathbb{E}A_{1;\tau})^{\tau^{-1}}\hat{\lambda}^*_{\tau;m}$ is asymptotically unbiased, establishing Part (1). Part (2) follows from Part (1) and Eqn (3). Finally, observe that $\mathbb{V}(\hat{\lambda}_{\tau;m}) = (\mathbb{E}A_{1;\tau})^{2\tau^{-1}}\mathbb{V}(\hat{\lambda}^*_{\tau;m})$. Since $\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}^*_{\tau;m}) = \tau^{-2}(\mathbb{E}A_{1;\tau})^{-2\tau^{-1}-2}\mathbb{V}(A_{1;\tau})$, we have

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \tau^{-2}(\mathbb{E}A_{1;\tau})^{-2}\mathbb{V}(A_{1;\tau}),$$

proving Part (3). □

Theorems 1 and 2 give us a simple recipe for calculating the limiting relative variance of τ-GRA-based estimators. In Sections 4 and 5 we follow the following three-step process:

1. Calculate the mean $\mu = \mathbb{E}A_{1;\tau}$ and the variance $\sigma^2 = \mathbb{V}(A_{1;\tau})$ of the τ-generalized remaining area at density 1.

2. By Theorem 1, the induced estimator $\hat{\lambda}^*_{\tau;m} = \left(\frac{1}{m}\sum_{i=1}^m A^{(i)}_{\lambda;\tau}\right)^{-\tau^{-1}}$ is a scale-invariant estimator for λ, but possibly biased.

3. After normalization, we get the estimator $\hat{\lambda}_{\tau;m} = \mu^{\tau^{-1}}\hat{\lambda}^*_{\tau;m}$ which is asymptotically unbiased. By Theorem 2, its relative variance is asymptotically $\tau^{-2}\mu^{-2}\sigma^2/m$.

4 Generalized Remaining Area for the **LogLog** Sketch

Consider a LogLog sketch consisting of m subsketches (columns in Figure 1), and let us focus on one subsketch with offset R. At cardinality λ, let X_λ be the index (an integer) of the highest occupied cell in this subsketch. Recall that the τ-generalized remaining area for this subsketch is summed up cell-by-cell:

$$\sum_{i=X_\lambda+1}^{\infty} (2^{-i-R})^\tau = \frac{1}{2^\tau-1}2^{-\tau(R+X_\lambda)} \propto 2^{-\tau(R+X_\lambda)}.$$

Because the cell sizes decay geometrically, this is linearly equivalent to taking the remaining area of the whole subsketch, $2^{-(R+X_\lambda)}$, to the τth power. For simplicity we calculate τ-GRA in this way, summing over subsketches rather than cells, thereby avoiding the leading constant $1/(2^\tau-1)$. Thus, $A_{\lambda;\tau} = 2^{-\tau(R+X_\lambda)}$ is the contribution of this subsketch to the τ-GRA.

Now we analyze X_λ. Fix an offset $r \in [0,1)$. For any $x > 0$, the event that $X_\lambda \le x$ is the event that the cells at or above $\lfloor x \rfloor + 1$ are all free. The sum of the heights of those cells is equal to $2^{-(\lfloor x \rfloor + r)}$ and they all have width $1/m$. Thus at cardinality λ, the number of darts in those cells is Poisson($m^{-1}\lambda 2^{-\lfloor x \rfloor + r}$). We have

$$\mathbb{P}(X_\lambda \le x|R=r) = e^{-m^{-1}\lambda 2^{-(\lfloor x \rfloor + r)}}.$$

We then characterize the distribution of $A_{\lambda;\tau} = 2^{-\tau(R+X_\lambda)}$ by the following lemma.

Lemma 1. *For any $x > 0$,*

$$\mathbb{P}(A_{\lambda;\tau} \ge x) = \int_0^1 e^{-x^{\tau^{-1}}2^r m^{-1}\lambda}\,dr.$$

12

Proof. Let $x > 0$ and $r \in [0, 1)$. Then we have

$$\mathbb{P}(A_{\lambda;\tau} \geq x | R = r) = \mathbb{P}(X_\lambda \leq -\log_2(x^{\tau^{-1}}) - r | R = r) = \exp(-m^{-1}\lambda 2^{-(\lfloor -\log_2(x^{\tau^{-1}}) - r \rfloor + r)}).$$

Set $y = \lfloor -\log_2(x^{\tau^{-1}}) - r \rfloor + r$ and $k = \lfloor -\log_2(x^{\tau^{-1}}) \rfloor$. Then for $r \in [0, -\log_2(x^{\tau^{-1}}) - k]$, we have $y = k + r$. For $r \in (-\log_2(x^{\tau^{-1}}) - k, 1)$, we have $y = k - 1 + r$. Therefore y iterates $(-\log_2(x^{\tau^{-1}}) - 1, -\log_2(x^{\tau^{-1}})]$ as r iterates $[0, 1)$. Therefore, we have

$$\mathbb{P}(A_{\lambda;\tau} \geq x) = \int_0^1 \mathbb{P}(A_{\lambda;\tau} \geq x | R = r) \, dr = \int_0^1 \exp(-m^{-1}\lambda 2^{-(\lfloor -\log_2(x^{\tau^{-1}}) - r \rfloor + r)}) \, dr$$

$$= \int_0^1 \exp(-m^{-1}\lambda 2^{\log_2(x^{\tau^{-1}}) + r}) \, dr = \int_0^1 e^{-x^{\tau^{-1}} 2^r m^{-1}\lambda} \, dr.$$

$\square$

We now prove that $A_{\lambda;\tau}$ is τ-scale-invariant.

Lemma 2. *For any $\tau > 0$, $A_{\lambda;\tau}$ is a τ-scale-invariant estimator for $\lambda^{-\tau}$.*

Proof. We need to prove that, for any $\tau, \lambda > 0$, $A_{\lambda;\tau} \sim \lambda^{-\tau} A_{1;\tau}$. Now note that, for any $x > 0$, we have

$$\mathbb{P}(A_{\lambda;\tau} \geq x) = \int_0^1 e^{-x^{\tau^{-1}} 2^r m^{-1}\lambda} \, dr = \int_0^1 e^{-(\lambda x^{\tau^{-1}})2^r \cdot m^{-1}} \, dr$$

$$= \mathbb{P}(A_{1;\tau} \geq \lambda^\tau x) = \mathbb{P}(\lambda^{-\tau} A_{1;\tau} \geq x).$$

$\square$

Recall that Γ is the continuous extension of the factorial function, with $\Gamma(n + 1) = n!$ when $n \in \mathbb{N}$. Its integral form is $\Gamma(z) = \int_0^\infty u^{z-1} e^{-u} du$.

Proposition 1. *For any $\tau > 0$,*

$$m^{-\tau} \mathbb{E}A_{1;\tau} = \Gamma(\tau)\frac{1 - 2^{-\tau}}{\log 2}, \quad and \quad m^{-2\tau} \mathbb{V}(A_{1;\tau}) = \Gamma(2\tau)\frac{1 - 2^{-2\tau}}{\log 2} - \left(\Gamma(\tau)\frac{1 - 2^{-\tau}}{\log 2}\right)^2.$$

Proof. We first prove some useful identities. Let $a, b > 0$. Then

$$\eta(a, b) \overset{\mathbf{def}}{=} \int_0^\infty e^{-ax^b} \, dx = \int_0^\infty e^{-t} a^{-1} b^{-1} (t/a)^{-\frac{b-1}{b}} \, dt = a^{-b^{-1}} b^{-1} \Gamma(b^{-1}),$$

where we set $t = ax^b$. Let $q, c > 0$.

$$\xi(q, b, c) \overset{\mathbf{def}}{=} \int_0^\infty \int_0^1 e^{-q^r x^b c} \, dr dx = \int_0^1 \eta(cq^r, b) \, dr = c^{-b^{-1}} b^{-1} \Gamma(b^{-1}) \int_0^1 q^{-rb^{-1}} \, dr$$

$$= c^{-b^{-1}} b^{-1} \Gamma(b^{-1})\frac{1 - q^{-b^{-1}}}{b^{-1} \log q} = c^{-b^{-1}} \Gamma(b^{-1})\frac{1 - q^{-b^{-1}}}{\log q}.$$

13

The first and second moments can now be calculated as follows.

$$\mathbb{E}A_{1;\tau} = \int_0^\infty \mathbb{P}(A_{1;\tau} \geq x)\,dx = \int_0^\infty \int_0^1 e^{-2^r x^{\tau^{-1}} m^{-1}}\,dr dx$$

$$= \xi(2, \tau^{-1}, m^{-1}) = m^\tau \Gamma(\tau) \frac{1 - 2^{-\tau}}{\log 2}$$

Turning to the second moment,

$$\mathbb{E}A_{1;\tau}^2 = \int_0^\infty \mathbb{P}(A_{1;\tau}^2 \geq x)\,dx = \int_0^\infty \mathbb{P}(A_{1;\tau} \geq x^{1/2})\,dx$$

$$= \int_0^\infty \int_0^1 e^{-2^r x^{\tau^{-1}/2} m^{-1}}\,dr dx = \xi\left(2, \frac{\tau^{-1}}{2}, m^{-1}\right)$$

$$= m^{2\tau} \Gamma(2\tau) \frac{1 - 2^{-2\tau}}{\log 2}.$$

We obtain the following closed form expression of the variance.

$$\mathbb{V}(A_{1;\tau}) = \mathbb{E}A_{1;\tau}^2 - (\mathbb{E}A_{1;\tau})^2 = m^{2\tau}\Gamma(2\tau)\frac{1 - 2^{-2\tau}}{\log 2} - m^{2\tau}\left(\Gamma(\tau)\frac{1 - 2^{-\tau}}{\log 2}\right)^2.$$

$\square$

Theorem 3 (τ-GRA for the LogLog sketch). *Let the offset vector $(R_i) \in [0,1)^m$ be selected uniformly at random. Let $X_\lambda^{(i)}$ be the integer index of the highest one in the ith subsketch after λ insertions. Then for any $\tau > 0$,*

$$\hat{\lambda}_{\tau;m} = m\left(\Gamma(\tau)\frac{1 - 2^{-\tau}}{\log 2}\right)^{\tau^{-1}}\left(\frac{1}{m}\sum_{i=1}^m 2^{-\tau(R_i + X_\lambda^{(i)})}\right)^{-\tau^{-1}}$$

is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is

$$\lim_{m \to \infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \tau^{-2}\left(\frac{\Gamma(2\tau)\log 2}{\Gamma(\tau)^2} \cdot \frac{1 + 2^{-\tau}}{1 - 2^{-\tau}} - 1\right).$$

Proof. This follows directly from Theorem 2 and Proposition 1. $\square$

Remark 2. The celebrated estimator $\hat{\lambda}_{\text{FFGM}}$ of HyperLogLog corresponds to $\tau = 1$. Inserting $\tau = 1$ to the variance formula, we have $\Gamma(2) = \Gamma(1) = 1$ and the leading constant of the variance is $3\log 2 - 1 \approx 1.07944$. The bias term at $\tau = 1$ is $\frac{1}{2\log 2}$, which match the constants from Flajolet et al. [FFGM07] as $m \to \infty$.

Remark 3. Note that for any $x_1, x_2, \ldots, x_m > 0$, $\lim_{\tau \to 0}\left(\frac{1}{m}\sum_{i=1}^m x_i^{-\tau}\right)^{-\tau^{-1}} = \left(\prod_{i=1}^m x_i\right)^{m^{-1}}$, i.e., the τ-mean converges towards the geometric mean as $\tau \to 0$. In other words, Durand and Flajolet's estimator $\hat{\lambda}_{\text{DF}}$ for LogLog corresponds to 0-GRA. We have the normalized relative variance[11]

$$\lim_{\tau \to 0} \tau^{-2}\left(\frac{\Gamma(2\tau)\log 2}{\Gamma(\tau)^2} \cdot \frac{1 + 2^{-\tau}}{1 - 2^{-\tau}} - 1\right) = \frac{2\pi^2 + \log^2 2}{12} \approx 1.68497,$$

which matches the limiting constant calculated by Durand and Flajolet [DF03, Dur04].

---

[11] This calculation is done in the algebraic system *Mathematica*.

See Figure 2 as a visualization of the relative variance of the τ-GRA estimators for the LogLog sketch. By numerical optimization, the minimal variance 1.07507 is obtained at $\tau^* = 0.889897$. This comes quite close to the Cramér-Rao lower bound for LogLog sketches, which Pettie and Wang [PW21] computed to be $\frac{\log 2}{\pi^2/6-1} \approx 1.07475$.

5   Generalized Remaining Area for the PCSA Sketch

Consider a PCSA sketch with m subsketches. Due to Poissonization, the sketch consists of a set of independent indicator variables corresponding to whether each cell has been hit by at least one dart. To simplify notation, let $X(t)$ be a "fresh" binary random variable such that

$$X(t) = \begin{cases} 0, & \text{with probability } e^{-t} \\ 1, & \text{with probability } 1 - e^{-t}. \end{cases}$$

Consider one subsketch with offset R. Cell i has height $2^{-(i+R)}$ and width $1/m$. At cardinality λ the number of points in the cell is $\text{Poisson}(m^{-1}\lambda 2^{-(i+R)})$. Thus the bit vector representing this subsketch distributes identically with $(X(m^{-1}\lambda 2^{-(i+R)}))_{i\in\mathbb{Z}}$. Let $\mathbb{1}\{\mathcal{E}\}$ denote the indicator variable for event $\mathcal{E}$. The τ-generalized remaining area for the PCSA sketch is then defined as follows.

$$A_{\lambda;\tau} = \sum_{i\in\mathbb{Z}} \mathbb{1}\left\{X(m^{-1}\lambda 2^{-(i+R)}) = 0\right\} 2^{-(i+R)\tau}.$$

Lemma 3. *For any $\tau > 0$, $A_{\lambda;\tau}$ is a τ-scale-invariant estimator for $\lambda^{-\tau}$.*

Proof. We need to prove that for any $\lambda > 0$, $A_{\lambda;\tau} \sim \lambda^{-\tau}A_{1;\tau}$. Note that

$$A_{\lambda;\tau} = \sum_{i\in\mathbb{Z}} \mathbb{1}\left\{X(m^{-1}\lambda 2^{-(i+R)}) = 0\right\} 2^{-\tau(i+R)}$$

$$= \lambda^{-\tau} \sum_{i\in\mathbb{Z}} \mathbb{1}\left\{X(m^{-1}2^{-(i+R-\log_2\lambda)}) = 0\right\} 2^{-\tau(i+R-\log_2\lambda)}$$

Note that because R is uniform over $[0,1)$ and we are summing over $\mathbb{Z}$, this sum is invariant under shifts, e.g., by $\log_2\lambda$. Continuing,

$$A_{\lambda;\tau} \sim \lambda^{-\tau} \sum_{i\in\mathbb{Z}} \mathbb{1}\left\{X(m^{-1}2^{-(i+R)}) = 0\right\} 2^{-\tau(i+R)} = \lambda^{-\tau}A_{1;\tau}.$$

$\square$

In contrast to our smoothing of (Hyper)LogLog, it actually *does* matter that we use the uniform offset vector $\mathbf{R} = (0, 1/m, \ldots, (m-1)/m)$ rather than random offsets. Random offsets would introduce subtle correlations between cells in the same column, and increase the variance by some tiny constant. Uniform offsets have the property that there is a cell of size $2^{-i/m}$ for every $i \in \mathbb{Z}$, so the conceptual organization of cells into columns is no longer relevant.

Proposition 2. *Let $A_{1;\tau}^{(1)}, A_{1;\tau}^{(2)}, \ldots, A_{1;\tau}^{(m)}$ be the τ-GRA of m subsketches with uniform offsetting. For $\tau > 0$,*

$$\lim_{m\to\infty} m^{-1-\tau} \sum_{i=1}^{m} \mathbb{E}(A_{1;\tau}^{(i)}) = \frac{\Gamma(\tau)}{\log 2}, \quad and \quad \lim_{m\to\infty} m^{-1-2\tau} \sum_{i=1}^{m} \mathbb{V}(A_{1;\tau}^{(i)}) = \frac{(1 - 2^{-2\tau})\Gamma(2\tau)}{\log 2}.$$

Proof. First note the following identity. Assume $q > 1, \tau > 0, \lambda > 0$. Then

$$\psi(\lambda, q, \tau) = \int_{-\infty}^{\infty} e^{-q^{-x}\lambda} q^{-\tau x} \, dx = \int_{0}^{\infty} e^{-t}(t/\lambda)^{\tau}(t \log q)^{-1} \, dt = \frac{1}{\log q} \lambda^{-\tau}\Gamma(\tau).$$

Here $t = q^{-x}$. After uniform offsetting, a PCSA sketch with m subsketches have cells of size $2^{-i/m}$ for all $i \in \mathbb{Z}$. Thus we we have

$$\lim_{m\to\infty} m^{-1-\tau} \sum_{i=1}^{m} \mathbb{E}\left(A_{1;\tau}^{(i)}\right) = \lim_{m\to\infty} m^{-1} \sum_{i\in\mathbb{Z}} \mathbb{E}\left(\mathbb{1}\left\{X(m^{-1}2^{-i/m}) = 0\right\}(m^{-1}2^{-(i/m)})^{\tau}\right)$$

Setting $h(t) = \mathbb{E}\left(\mathbb{1}\left\{X(2^{-t}) = 0\right\}(2^{-t})^{\tau}\right) = e^{-2^{-t}}2^{-\tau t}$, the sum becomes $m^{-1}\sum_{i\in\mathbb{Z}} h(i/m+\log_2 m)$. Since we are summing over $\mathbb{Z}$, the shift $\log_2 m$ in the argument affects the sum vanishingly as $m \to \infty$. Thus $\lim_{m\to\infty} m^{-1}\sum_{i\in\mathbb{Z}} h(i/m + \log_2 m) = \lim_{m\to\infty} m^{-1}\sum_{i\in\mathbb{Z}} h(i/m) = \int_{-\infty}^{\infty} h(t)\,dt$. Thus,

$$\lim_{m\to\infty} m^{-1-\tau} \sum_{i=1}^{m} \mathbb{E}\left(A_{1;\tau}^{(i)}\right) = \int_{-\infty}^{\infty} e^{-2^{-t}}2^{-\tau t}\,dt = \psi(1, 2, \tau) = \frac{\Gamma(\tau)}{\log 2}.$$

Note that by Poissonization, cells are independent and thus all co-variances are zero.

$$\lim_{m\to\infty} m^{-1-2\tau} \sum_{i=1}^{m} \mathbb{V}(A_{1;\tau}^{(i)}) = \lim_{m\to\infty} m^{-1} \sum_{i\in\mathbb{Z}} \mathbb{V}\left(\mathbb{1}\left\{X(m^{-1}2^{-i/m}) = 0\right\}(m^{-1}2^{-i/m})^{\tau}\right)$$

by the same limiting argument laid out above, this is equal to

$$= \int_{-\infty}^{\infty} \mathbb{V}(\mathbb{1}\left\{X(2^{-t}) = 0\right\}2^{-\tau t})\,dt$$

$$= \int_{-\infty}^{\infty} e^{-2^{-t}}2^{-2\tau t} - e^{-2\cdot2^{-t}}2^{-2\tau t}\,dt$$

$$= \psi(1, 2, 2\tau) - \psi(2, 2, 2\tau) = \frac{\Gamma(2\tau)}{\log 2}(1 - 2^{-2\tau}).$$

$\square$

Theorem 4 (τ-GRA for the PCSA sketch)**.** *Let $A_{\lambda;\tau}^{(i)}$ be the τ-generalized remaining area of the ith subsketch with uniform offsetting, and $A = \sum_{i=1}^{m} A_{\lambda;\tau}^{(i)}$ be the τ-GRA. Then for any $\tau > 0$,*

$$\hat{\lambda}_{\tau;m} = m\left(\frac{\Gamma(\tau)}{\log 2}\right)^{\tau^{-1}}\left(\frac{A}{m}\right)^{-\tau^{-1}}$$

is a scale-invariant estimator for λ that is asymptotically unbiased. The asymptotic normalized relative variance is

$$\lim_{m\to\infty} m\lambda^{-2}\mathbb{V}(\hat{\lambda}_{\tau;m}) = \frac{(1 - 2^{-2\tau})\Gamma(2\tau)\log 2}{\tau^2\Gamma(\tau)^2}.$$

16

Proof. This follows directly from Theorem 2 and Proposition 2. $\qquad\square$

Remark 4. The remaining area estimator (1-GRA) has normalized relative variance $\frac{(1-2^{-2})\Gamma(2)}{1^2\Gamma(1)^2}\log 2 = \frac{3}{4}\log 2 \approx 0.51986$, which is better than Flajolet and Martin's original "first zero" estimator $\hat{\lambda}_{\text{FM}}$.

Remark 5. As τ goes to 0, $\hat{\lambda}_{\tau;m}$ is essentially counting the number of free cells (0s in the sketch), which corresponds to Lang's [Lan17] "coupon collector" estimator $\hat{\lambda}_{\text{Lang}}$ that counts occupied cells (1s in the sketch). The limiting variance of this estimator is[12]

$$\lim_{\tau\to 0}\frac{(1-2^{-2\tau})\Gamma(2\tau)\log 2}{\tau^2\Gamma(\tau)^2} = \log^2 2 \approx 0.480453,$$

which confirms Lang's [Lan17] back-of-the-envelope calculation that it should be $\log^2 2$.

See Figure 3 for a visualization of the relative variance of the τ-GRA estimators for the PCSA sketch. By numerical optimization, the minimal variance 0.435532 is obtained at $\tau^* = 0.343557$. This comes very close to the Cramér-Rao lower bound of $\frac{\pi^2}{6\log 2} \approx 0.42138$ for PCSA sketches, as computed in [PW21].

6   Conclusion

We introduced a class of estimators based on the concept of τ-*generalized remaining area*, which has significant explanatory power, as it subsumes many existing estimators [DF03, FFGM07, Lan17]. This concept is quite powerful, and allows us to *almost* completely close the gap between the best explicit (non-MLE) estimators for HyperLogLog [FFGM07] and PCSA [Lan17] and their respective Cramér-Rao lower bounds [PW21]. See Figures 2 and 3 and Table 1.

One distinction of our proofs is that they are very precise, but assume only basic probability and calculus. They avoid the daunting complexity of a Flajolet-style analysis [FM85, Fla90, DF03, FFGM07]. The key ingredients in our approach are (i) a deliberate simplification of the probabilistic model (see Definitions 1, 2, and 3), and (ii) restricting the analysis to *limiting* relative variance rather than try to understand the variance at *every* value of m.[13]

We hope that our analysis will make the popular and efficient cardinality sketches accessible to students, at least at the graduate level. At present, courses on Big Data/Subliner Algorithms [Pri20, Che14, Yar15, Cha20, Vas11, Nel20, McG18, Mus21, Woo20, Mah21, IR13] usually cover Cardinality Estimation/Distinct Elements, but avoid HyperLogLog, PCSA, and related sketches in favor of sketches with simpler analyses.

References

[AMS99]   Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *J. Comput. Syst. Sci.*, 58(1):137–147, 1999.

---

[12]This calculation is done in the algebraic system *Mathematica*.

[13]In practice m is typically between 2^8 and 2^{14}, which is already in the asymptotic regime. A Flajolet-style analysis—which obtains bias correction constants and a precise variance when m is *fixed*—is most helpful when m is a small constant.

[BGH+09] Kevin S. Beyer, Rainer Gemulla, Peter J. Haas, Berthold Reinwald, and Yannis Sismanis. Distinct-value synopses for multiset operations. *Commun. ACM*, 52(10):87–95, 2009.

[BJK+02] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, D. Sivakumar, and Luca Trevisan. Counting distinct elements in a data stream. In *Proceedings 6th International Workshop on Randomization and Approximation Techniques (RANDOM)*, volume 2483 of *Lecture Notes in Computer Science*, pages 1–10, 2002.

[BKS02] Ziv Bar-Yossef, Ravi Kumar, and D. Sivakumar. Reductions in streaming algorithms, with an application to counting triangles in graphs. In *Proceedings 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 623–632, 2002.

[Bł20] Jarosław Błasiok. Optimal streaming and tracking distinct elements with high probability. *ACM Trans. Algorithms*, 16(1):3:1–3:28, 2020.

[CB02] G. Casella and R. L. Berger. *Statistical Inference, 2nd Ed.* Brooks/Cole, Belmont, CA, 2002.

[CCSN11] Aiyou Chen, Jin Cao, Larry Shepp, and Tuan Nguyen. Distinct counting with a self-learning bitmap. *Journal of the American Statistical Association*, 106(495):879–890, 2011.

[CG06] Philippe Chassaing and Lucas Gerin. Efficient estimation of the cardinality of large data sets. In *Proceedings of the 4th Colloquium on Mathematics and Computer Science Algorithms, Trees, Combinatorics and Probabilities*, 2006.

[Cha20] A. Chakrabarti. CS 35/135: Data Stream Algorithms. `https://www.cs.dartmouth.edu/~ac/Teach/CS35-Spring20/`, 2020.

[Che14] C. Chekuri. CS 598: Algorithms for Big Data. `https://courses.engr.illinois.edu/cs598csc/fa2014/`, 2014.

[Coh97] Edith Cohen. Size-estimation framework with applications to transitive closure and reachability. *J. Comput. Syst. Sci.*, 55(3):441–453, 1997.

[Coh15] Edith Cohen. All-distances sketches, revisited: HIP estimators for massive graphs analysis. *IEEE Trans. Knowl. Data Eng.*, 27(9):2320–2334, 2015.

[DF03] Marianne Durand and Philippe Flajolet. Loglog counting of large cardinalities. In *Proceedings 11th Annual European Symposium on Algorithms (ESA)*, volume 2832 of *Lecture Notes in Computer Science*, pages 605–617. Springer, 2003.

[Dur04] Marianne Durand. *Combinatoire analytique et algorithmique des ensembles de données. (Multivariate holonomy, applications in combinatorics, and analysis of algorithms).* PhD thesis, Ecole Polytechnique X, 2004.

[Ert17] Otmar Ertl. New cardinality estimation methods for HyperLogLog sketches. *CoRR*, abs/1706.07290, 2017.

[EVF06]   Cristian Estan, George Varghese, and Michael E. Fisk. Bitmap algorithms for counting active flows on high-speed links. *IEEE/ACM Trans. Netw.*, 14(5):925–937, 2006.

[FFGM07] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In *Proceedings of the 18th International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods for the Analysis of Algorithms (AofA)*, pages 127–146, 2007.

[Fla90]   Philippe Flajolet. On adaptive sampling. *Computing*, 43(4):391–400, 1990.

[FM85]   Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *J. Comput. Syst. Sci.*, 31(2):182–209, 1985.

[Gir09]   Frédéric Giroire. Order statistics and estimating cardinalities of massive data sets. *Discret. Appl. Math.*, 157(2):406–427, 2009.

[GT01]   Phillip B. Gibbons and Srikanta Tirthapura. Estimating simple functions on the union of data streams. In *Proceedings 13th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA)*, pages 281–291, 2001.

[HNH13]   Stefan Heule, Marc Nunkesser, and Alexander Hall. HyperLogLog in practice: algorithmic engineering of a state of the art cardinality estimation algorithm. In *Proceedings 16th International Conference on Extending Database Technology (EDBT)*, pages 683–692, 2013.

[IR13]   P. Indyk and R. Rubinfeld. 6.893: Sub-linear Algorithms. `http://stellar.mit.edu/S/course/6/sp13/6.893/`, 2013.

[IW03]   Piotr Indyk and David P. Woodruff. Tight lower bounds for the distinct elements problem. In *Proceedings 44th IEEE Symposium on Foundations of Computer Science (FOCS), October 2003, Cambridge, MA, USA, Proceedings*, pages 283–288, 2003.

[JW13]   T. S. Jayram and David P. Woodruff. Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error. *ACM Trans. Algorithms*, 9(3):26:1–26:17, 2013.

[KNW10]   Daniel M. Kane, Jelani Nelson, and David P. Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings 29th ACM Symposium on Principles of Database Systems (PODS)*, pages 41–52, 2010.

[KP22]   Matti Karppa and Rasmus Pagh. HyperLogLogLog: Cardinality estimation with one log more. In *Proceedings 28th ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 753–761, 2022.

[Lan17]   Kevin J. Lang. Back to the future: an even more nearly optimal cardinality estimation algorithm. *CoRR*, abs/1708.06839, 2017.

[ŁU22]   Aleksander Łukasiewicz and Przemysław Uznański. Cardinality estimation using Gumbel distribution. In *Proceedings 30th European Symposium on Algorithms (ESA)*, 2022.

[Lum10]   Jérémie Lumbroso. An optimal cardinality estimation algorithm based on order statistics and its full analysis. In *Proceedings of the 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA)*, pages 489–504, 2010.

[Mah21]   S. Mahabadi. TTIC 41000: Algorithms for Massive Data. `https://www.mit.edu/~mahabadi/courses/Algorithms_for_Massive_Data_SP21/`, 2021.

[McG18]   A. McGregor. CMPSCI 711: More Advanced Algorithms. `https://people.cs.umass.edu/~mcgregor/CS711S18/`, 2018.

[Mus21]   C. Musco. COMPSCI 514: Algorithms for Data Science. `https://people.cs.umass.edu/~cmusco/CS514F21/`, 2021.

[Nel20]   J. Nelson. CS 294-165: Sketching Algorithms. `https://www.sketchingbigdata.org/fall20/`, 2020.

[Oha21]   Tal Ohayon. ExtendedHyperLogLog: Analysis of a new cardinality estimator. *CoRR*, abs/2106.06525, 2021.

[Pri20]   E. Price. CS395T: Sublinear Algorithms. `https://www.cs.utexas.edu/~ecprice/courses/sublinear/`, 2020.

[PW21]   Seth Pettie and Dingyu Wang. Information theoretic limits of cardinality estimation: Fisher meets Shannon. In *Proceedings 53rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 556–569, 2021.

[PWY21]   Seth Pettie, Dingyu Wang, and Longhui Yin. Non-mergeable sketching for cardinality estimation. In *Proceedings 48th International Colloquium on Automata, Languages, and Programming (ICALP)*, volume 198 of *LIPIcs*, pages 104:1–104:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[SM07]   Björn Scheuermann and Martin Mauve. Near-optimal compression of probabilistic counting sketches for networking applications. In *Proceedings of the 4th International Workshop on Foundations of Mobile Computing (DIALM-POMC)*, 2007.

[The19]   The Apache Foundation. Apache DataSketches: A software library of stochastic streaming algorithms. `https://datasketches.apache.org/`. 2019.

[Tin14]   Daniel Ting. Streamed approximate counting of distinct elements: beating optimal batch methods. In *Proceedings 20th ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 442–451, 2014.

[Vaa98]   A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[Vas11]   S. Vassilvitskii. COMS 6998-12: Dealing with Massive Data. `http://www.cs.columbia.edu/~coms699812/`, 2011.

[Woo20]   D. Woodruff. 15-859: Algorithms for Big Data. `https://www.cs.cmu.edu/~dwoodruf/teaching/15859-fall20/`, 2020.

[XCZL20] Qingjun Xiao, Shigang Chen, You Zhou, and Junzhou Luo. Estimating cardinality for arbitrarily large data stream with improved memory efficiency. *IEEE/ACM Trans. Netw.*, 28(2):433–446, 2020.

[Yar15] G. Yaroslavtsev. CIS 700: algorithms for Big Data. `http://grigory.us/big-data-class.html`, 2015.