Approximate Distinct Counts for Billions of Datasets

Daniel Ting
Tableau Software
Seattle, Washington
dting@tableau.com

ABSTRACT

Cardinality estimation plays an important role in processing big data. We consider the challenging problem of computing millions or more distinct count aggregations in a single pass and allowing these aggregations to be further combined into coarser aggregations. These arise naturally in many applications including networking, databases, and real-time business reporting. We demonstrate existing approaches to solve this problem are inherently flawed, exhibiting bias that can be arbitrarily large, and propose new methods for solving this problem that have theoretical guarantees of correctness and tight, practical error estimates.

This is achieved by carefully combining CountMin and HyperLogLog sketches and a theoretical analysis using statistical estimation techniques. These methods also advance cardinality estimation for individual multisets, as they provide a provably consistent estimator and tight confidence intervals that have exactly the correct asymptotic coverage.

KEYWORDS

Distinct Counting, Cardinality estimation, HyperLogLog, CountMin, Approximate Query Processing

ACM Reference Format:
Daniel Ting. 2019. Approximate Distinct Counts for Billions of Datasets. In *2019 International Conference on Management of Data (SIGMOD '19), June 30-July 5, 2019, Amsterdam, Netherlands.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3299869.3319897

1 INTRODUCTION

Distinct count aggregations are particularly important but costly statistics for database applications. As a result, approximate distinct counting sketches have enjoyed widespread adoption in databases and in other applications. These enable arbitrarily large distinct counts to be computed in a single pass with just a few kilobytes and errors under 1%. However, a number of applications require instantiating many counters, sometimes in the billions and beyond. In such cases, even a few kilobytes for every counter is too much.

These cases arise whenever aggregate statistics must be computed for a large number of subgroups or datasets. For example, networking applications may compute the number of distinct flows per destination IP [16] to detect DDoS attacks. A web search engine may compute the number of distinct users that have used a word in a query or clicked on a search result, or web advertisers may wish to know how many distinct users have seen each of their ads [24]. The use cases include a broad range of applications and are not restricted to reporting. For example, [2] estimate the diameter of a large social network by computing the number of distinct users d relationships away. In each case, billions of distinct counts are computed with cardinalities that can also reach a billion and beyond. These cases can occur for non-web scale data as well. A group-by clause with multiple categorical columns can generate exponentially many counters in the number of columns. Similarly, in OLAP cube applications, the number of vertices of the cube grows exponentially as more possible breakdowns are admitted.

Although the problem of approximating a single distinct count is well-studied, few methods [9, 35] do so when the number of distinct counters explodes. We show these methods can suffer severe deficiencies in performance, and these deficiencies appear when applied to real data sets. Furthermore, there are limited error guarantees and no way to detect when the performance is poor. This makes it difficult or impossible to build systems where engineers can be guaranteed that performance will be acceptable for their workloads and users can be assured that the results can be trusted.

This work addresses these deficiencies, yielding sketches and distinct count estimators which dominate previous ones in accuracy while still maintaining a fixed size and $O(1)$ updates. For each, we obtain guarantees that the approximated counts converge to the truth, and derive tight confidence intervals that deliver almost exactly the promised coverage.

We also consider the problem of merging counters, both across sketches and within a single sketch. This allows the functionality of the sketch to be close to that of storing individual distinct counting sketches. We discuss practical

SIGMOD '19, June 30-July 5, 2019, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5643-5/19/06.
https://doi.org/10.1145/3299869.3319897

implementation issues including the choice of sketch parameters and tradeoffs in terms of functionality, theoretical guarantees, and reduction of noise. Our work also applies to cardinality estimation on a single dataset. It yields a new state-of-the-art estimator that requires no empirical bias correction and is the first to provide tight error estimates over the entire range of cardinalities.

1.1 Related Work

Past work has extensively studied the problem of approximately computing the cardinality, or number of distinct items, in a single dataset as well as the problem of approximately computing a non-distinct count for many multisets keyed by some label. For estimating a single cardinality, many methods exist such as Probabilistic Counting (PCSA) [18], Linear Probabilistic Counting [34], HyperLogLog (HLL) [17], Multiresolution Bitmap [16]. While there continues to be research on improving these sketches [7, 15, 19, 21, 29], the possible gains in improving single multiset cardinality estimators appear limited. The optimal space complexity has been achieved [20]. Other theoretical results give optimal estimators with the best possible constants for any sketch [29], and information theoretic results provide lower bounds on the space occupied by a sketch construction [21]. Thus, there is little room for improvement when estimating large counts. For small counts, approaches using sparse representations [19] can significantly reduce space usage.

When there are many multisets keyed by some label, CountMin [12] and the Alon-Gibbons-Matias-Szegedy (AGMS) sketch [1, 11, 28] address the problem of approximately computing non-distinct counts for each label. Like the distinct counting results, [31] provide a provably optimal estimator for the CountMin sketch. If only large counts are of interest, frequent item sketches such as Misra-Gries [26], Spacesaving [25], Lossy counting [23], and RAP [3] can provide both a estimated list of frequent items and counts.

These counting sketches can provide some ability to perform further aggregation as well since any counting sketch can approximate the count over multiple labels as long as the labels can be enumerated. If the labels cannot be enumerated, for example when the aggregation is over an unknown universe of labels that match some filter condition, Unbiased Space Saving [32] can be used.

This paper addresses the problem of estimating distinct counts for many multisets and allowing for further aggregations when the labels can be enumerated. We call this the many distinct count problem. Although a few past works [9, 35] have proposed solutions by combining distinct counting sketches with counting sketches, sections 6 and 7 show these past solutions are inefficient and can be highly biased.

Other approaches use coordinated or hash based sampling [8]. Since these can store actual samples, filtering, grouping, aggregation, and other set operations can be simple to perform [5, 30]. However, these sketches require an order of magnitude more space to achieve the same error as a HyperLogLog sketch for single multiset cardinality estimation. Furthermore, a cardinality estimate for a small subset still requires scanning the entire sketch and can result in high computational costs.

1.2 Many Distinct Count Problem

Consider a stream $\mathcal{D}$ of (item, label) pairs $(x_1, \ell_1), \ldots, (x_T, \ell_T)$ and the following queries that define the many distinct count problem.

- A pointwise count distinct query takes a label ℓ and returns the number n_ℓ of distinct items with label ℓ. We write $\mathcal{X}_\ell$ to be the set of items with label ℓ.
- An aggregate count distinct query takes a collection of labels $\ell_1, \ldots, \ell_c$ and returns $|\cup_{i=1}^{c} \mathcal{X}_{\ell_i}|$, the number of distinct items that match any label in the collection.
- The total count distinct query returns the total number of distinct items in the stream irrespective of the labels.

In terms of SQL, these span queries of the form

```
SELECT count(distinct x)
  FROM table
 WHERE label in (...)
```

where ... represents some list of labels or the case where the `in` clause is omitted entirely.

Denote by $\mathcal{L}$ the universe of distinct labels. We consider the regime where both the number of distinct items n_{tot} and the number of distinct labels $|\mathcal{L}|$ are large. We wish to construct a sketch C that has the following properties.

- C has bounded size.
- It provides approximate answers to all 3 types of queries.
- The estimators converges to the truth.
- It provides tight error estimates.
- Merging with sketch C' computes a union of datasets.

2 OVERVIEW

Although our paper addresses an important problem for databases, it employs statistical techniques that may not be familiar to the database researcher. We briefly introduce some of these statistical ideas. The confident reader may skip this section and revisit it as a summary. A table of symbols is provided in the appendix for reference.

Like other approximate distinct counting methods, our method relies on transforming a stream into a random process where the parameters of the distribution are the cardinalities. Cardinality estimation thus becomes a statistical parameter estimation problem where one wishes to infer the parameters θ that yield the true data generating distribution $p(Sketch|\theta)$. Unlike most statistical estimation problems

that require strong, unverifiable assumptions about the true model, in other words the functional form of $p(Sketch|\theta)$, the true model can be derived from the sketch construction.

Under this viewpoint, sketching algorithms can be split into two parts: construction, which is about designing the observable data i.e. the sketch, and estimation, which extracts information from the sketch by applying the appropriate statistical machinery. These can be further decomposed into the following steps.

- Sketch design } Construction
- Distribution identification } Construction
- Parameter estimation } Estimation
- Error estimation } Estimation

We cover these in reverse order since the design of good constructions is informed by a reducation of the estimation problem to a distribution identification problem. For those interested only in estimation for HLL sketches, section 3 on estimation is self-contained.

2.1 Estimation overview

Using existing statistical machinery for estimation provides several benefits. First, the methods typically come with theoretical guarantees. The most basic guarantee is consistency, or convergence to the truth. Second, they often provide tight error estimates. Third, many statistical methods are asymptotically efficient. They not only have an optimal error rate, but they attain the best possible constant governing that rate. Fourth, the methods are general. They can be applied to multiple sketch constructions with minor modifications.

A standard approach for statistical estimation is to apply maximum likelihood estimation. The likelihood function $L(\theta; Data) = p(Data|\theta)$ reinterprets the probability function as a function of the parameters rather than data. The data is fixed to be the observed data. The maximum likelihood estimate (MLE) is the cardinality under which the data has the highest probability of being observed. It plays a pivotal role in statistics as the likelihood principle [4] states that all information about the parameter is contained in the likelihood and the MLE is asymptotically efficient under some regularity conditions. Unfortunately, maximum likelihood cannot be applied in our cases as the likelihood function is intractable to compute. However, our procedure follows the same simple paradigm. Maximize a objective function measuring the goodness of an estimate. Base it on the true probability so that it is guaranteed to converge to the truth.

The particular method we use is composite marginal likelihood. This reduces the problem of estimation to that of computing the univariate, also known as marginal, distribution for each sketch entry. Summing over the observations gives the marginal likelihood and our objective function. Deriving these marginal distributions through operations on cumulative distribution functions (cdfs) and empirical cdfs is a major component of this paper. The cdf F of a random variable X is defined by $F(x) := P(X \leq x)$. It fully characterizes a distribution since $p(X = x) = F(x) - F(x-1)$ when X is integer valued, so we refer to a distribution interchangeably with its cdf. The empirical cdf is the cdf of the observed data, which is a random draw from the true underlying distribution.

2.2 Construction overview

We propose a pair of sketch constructions that combine the distinct counting abilities of HyperLogLog (HLL) with the counter compression abilities of CountMin. Both our and existing sketches are equivalent to first generating separate HLL sketches, one for each label, and then projecting them into a much smaller set of entries in a compressed sketch.

The differences among constructions are small. Every construction uses random hash functions and generates the same uniform distribution over registers and geometric distribution for the value for an (item, label) pair. However, by varying the inputs to the hash functions, they introduce different dependencies among data points. These small changes have significant impact on the properties of the sketch.

The main challenge of the construction phase is identifying the marginal distribution of sketch entries. This is needed for our composite likelihood estimator. The exact distribution is impossible to compute since it depends on the unknown cardinality of every label. The critical idea which enables us to identify the distribution is decomposing each sketch entry $C_i = S_i \wedge \epsilon_i$ into independent signal S_i and error terms ϵ_i with respect to a given label ℓ. Here $x \wedge y := \max\{x, y\}$ and the signal S_i corresponds to the i^{th} entry of a HLL sketch constructed just from items with label ℓ.

We can then circumvent the problem of an intractable exact error distribution by empirically estimating the independent error distribution from the sketch itself. Our estimators are non-parametric which means that there are no assumptions on the distribution of the cardinalities that might limit the resulting estimates' validity to a certain set of cases.

2.3 Asymptotics and assumptions

Although most of our theoretical results are asymptotic, they often accurately predict the finite sample behavior when the asymptotic regime mimics realistic scenarios. This is a general characteristic of many statistical estimation methods for problems with few parameters. The experiments in section 7 validate this. In contrast, guarantees derived from probability inequalities typically provide only a rate or an extremely loose bound which may not be of practical use.

The asymptotic regime we consider mimics the case where the sketch is large but the number of labels and items per

sketch entry is modest in size. More precisely, we consider the case where for sequence of sketches that are $d \times w$ arrays and some constants λ_ℓ

$$d \to \infty \tag{1}$$

$$n_\ell / d \to \lambda_\ell \quad \text{for all } \ell \in \mathcal{L}. \tag{2}$$

The number of labels $|\mathcal{L}|$ is either fixed or $|\mathcal{L}|/w \to \rho$ as $w \to \infty$ for some constant $\rho > 0$. When $|\mathcal{L}| \to \infty$, we assume the set for each label is an independent draw from some distribution Γ over sets. We will call the regime where $|\mathcal{L}|$ is fixed to be the standard asymptotic regime. When $|\mathcal{L}| \to \infty$, we refer to it as the width increasing regime. The regimes we consider avoid easy cases such as when the sketch grows faster than the number of items, as well as cases where one can exploit unrealistic asymptotic normality, such as the case when the number of labels per sketch entry goes to infinity.

In this regime, we are primarily concerned with consistency of the estimator $\hat{N}$. Specifically we show convergence in probability $\hat{N}_\ell / d \xrightarrow{p} \lambda_\ell$ to the truth. Equivalently, the relative error $(\hat{N}_\ell - n_\ell)/n_\ell \xrightarrow{p} 0$. We also give distributional convergence results that provide error estimates. These naturally imply that the estimators achieve the standard, and optimal, rate of convergence for parametric problems of $O(1/\sqrt{d})$.

Throughout, we assume that all hashes used are universal in the sense that for a random hash function h and for the data $\{X_i\}$, their hashed values $\{h(X_i)\}$ are a collection of independent, identically distributed (i.i.d.) random variables. This independence property may thus be a property of the hash function, of the data, or both.

3 ESTIMATION AND COMPOSITE-LIKELIHOOD

We first cover cardinality estimation using composite likelihood methods and use the HLL sketch [17] as an example. In particular, we use composite marginal likelihood which reduces the problem of cardinality estimation to that of computing the marginal probability of an individual sketch entry. While HLL has a particular parametric form for this marginal probability, the calculations in this section apply to other sketch constructions without modification. The only difference is that the marginal cumulative distribution function (cdf) for HLL entries is substituted with the cdf for the chosen construction.

This section is largely self contained. It also makes two contributions to estimation for basic HLL sketches. To the best of our knowledge, we provide the first provably consistent estimator of the cardinality. Furthemore, we provide the only tight error estimator and confidence intervals (CIs) for the entire range of cardinalities.

3.1 HyperLogLog sketch

We first introduce the HLL sketch and its construction. It is widely used in practice as it is accurate for wide range of cardinalities and highly space efficient, requiring < 8KB to approximate cardinalities exceeding 2^{50} with < 1% relative error [17].

The sketch consists of an array C of d small, non-negative integers which we refer to as registers. Each unique item x_t in the stream is hashed to a register $b_t = h(x_t)$ and value $z_t = g(x_t)$. Both h, g are random hashes where $h(x_i)$ is uniformly distributed and $g(x_i) \sim Geometric(1/2)$. The final value C_b of register b is the maximum of all values hashed to it. Succinctly, the random process generating the HLL sketch may be described as follows.

$$\begin{aligned} x_t &\overset{h}{\mapsto} B_t \sim Uniform(d) \\ x_t &\overset{g}{\mapsto} Z_t \sim Geometric(1/2) \\ C_b &= \max\{Z_i : B_i = b\} \end{aligned} \tag{3}$$

Here the notation $X \sim F$ denotes that the random variable X is drawn from the distribution F. In pseudocode this is represented by

function HLL-ADD(C, x)
  $b \leftarrow$ HASH$(x, seed_1)$ mod d
  $z \leftarrow$ TRAILINGZEROS(HASH$(x, seed_2)$) + 1
  $C[b] \leftarrow$ MAX$(C[b], z)$
end function

for some fixed seeds. Here, TrailingZeros refers to the number of trailing zeros in the binary representation of the hashed value. Since any duplicates x_i and x_j hash to exactly the same register and value, the sketch is unchanged by duplicates, making it suitable for distinct counting.

For large cardinalities, the original HLL estimator is $\hat{n}_{HLL} = d^2 \alpha_d \left(\sum_b 2^{-C_b}\right)^{-1}$ where α_d is some constant that depends on the sketch size d. Rather than asymptotic unbiasedness and consistency, the estimates are only guaranteed to have an asymptotic bias that is small [17]. They also provide an asymptotic error. These results are only for the simpler asymptotic regime where the number of items per register goes to infinity. Small to medium cardinalities require other estimators or empirical corrections [19]. The analysis of the estimates is quite involved, and it is unclear to us if this analytic approach can be applied to more general sketch constructions.

3.2 The max and HLL sketch distributions

To cast cardinality estimation as a statistical parameter estimation problem, we must derive the distribution of the HLL sketch and show it is a function of the cardinality. Equation 3 shows this distribution depends on the computing the maximum of independent random variables.

We use the standard convention of uppercase for random variables and lowercase for constant values. We also use the hat notation so that $\hat{\theta}$ denotes the estimate of θ. Consider independent random variables X, Y with cdfs $F_X(v) := P(X \leq v)$ and F_Y. Since the maximum $X \wedge Y \leq v$ if and only if both $X \leq v$ and $Y \leq v$, one has the following cdf

$$p(X \wedge Y \leq v) = F_X(v)F_Y(v) \tag{4}$$

In particular, if $\{X_i\}$ is a collection of i.i.d. random variables with cdf F, then $p(\bigwedge_{i=1}^{t} X_i \leq v) = F(v)^t$.

In the HLL construction for a set with n items, the items choose bins independently and uniformly at random. The total number K_b of items assigned to register C_b is multinomially distributed, and the register value distribution is given by the distribution of maximum given above. Succinctly,

$$K_1, \ldots, K_m \sim Multinomial(n_{tot}, 1/d, \ldots, 1/d) \tag{5}$$

$$C_b | K_b \sim G^{K_b} \tag{6}$$

where G is the cdf of a $Geometric(1/2)$ random variable. Here $C_b|K_b \sim G^{K_b}$ denotes that the conditional distribution of C_b given $K_b = k$ has cdf G^k. For the special case $K_b = 0$ we have $C_b = 0$. This shows that the only unknown parameter in the random process is the cardinality n_{tot}, since the sketch size d is known. The distribution of the sketch C can be written as $p(C|n)$ and derived from the equations above.

3.3 Composite-Likelihood Estimation

Identifying the data generating distribution enables the standard technique of maximum likelihood estimation to be applied. However, in this case, it is impractical since the likelihood is computationally intractable. The likelihood of the full HLL sketch is

$$\begin{aligned} p(C|n) &= \sum_K p(K)p(C|K) \\ &= \sum_{k_1+\cdots+k_d=n} \binom{n}{k_1, \ldots, k_d} \frac{1}{d}^n \prod_i (G(C_i)^{k_i} - G(C_i - 1)^{k_i}) \end{aligned}$$

The sum over integer partitions of n makes the likelihood difficult to compute.

Composite likelihood ($c\ell$) methods [22, 33] circumvent the problem by computing only a tractable portion of a full likelihood. The advantage of composite-likelihood techniques is that they remain asymptotically consistent under modest assumptions since they are still based on a true likelihood.

Specifically, we derive a composite marginal likelihood estimator. This estimator can be trivially generalized to work for all our sketch constructions. For the HLL sketch, it has state-of-the-art performance and improves upon previous work by providing theoretical consistency guarantees and tight error estimates for the entire range of cardinalities. Figure 1 confirms that our asymptotic theory is almost perfectly predictive even in finite sample settings. Our estimator has a Relative Root Mean Squared Error (RRMSE) curve that matches other state-of-the-art estimators [15] and our new estimated RRMSE almost perfectly matches the true RRMSE.

The marginal likelihood $p(C_b|n)$ is made computationally tractable by changing the representation so that each sketch entry is the maximum of a constant number of random variables. By treating items not hashed to bin b as 0, the register C_b can be alternatively expressed as

$$\begin{aligned} C_b &= \max\{\tilde{Z}_t : t \in 1, \ldots, n\} \\ \text{where } \tilde{Z}_t &= Z_t \text{ if } B_t = b \text{ else } 0 \end{aligned} \tag{7}$$

Applying equation 4 for the distribution of the maximum gives the cdfs for $\tilde{Z}_t, C_b$ and corresponding pmf for C_b.

$$\tilde{G}(v) := p(Z_t \leq v) = 1 - \frac{1}{m} + \frac{G(v)}{m} = 1 - \frac{1}{m2^v} \tag{8}$$

$$F(v|n) := p(C_b \leq v|n) = \tilde{G}(v)^n \tag{9}$$

$$f(v|n) = \tilde{G}(v)^n - \tilde{G}(v-1)^n \tag{10}$$

The composite marginal log-likelihood ($c\ell$) replaces the true log-likelihood with the sum of log marginal probabilities, and the maximum $c\ell$ estimator maximizes this approximation to the true log-likelihood.

$$c\ell(n; C) := \sum_b \log f(C_b|n) \tag{11}$$

$$\hat{N}_{c\ell} := \arg\max_n c\ell(n; C) \tag{12}$$

For implementation details, section 5 provides a Newton-based algorithm for a more general form of the composite likelihood needed for many distinct count sketches.

3.4 Consistency and error estimates

The $c\ell$ estimator has attractive theoretical properties. The main property retained by composite likelihood estimates is consistency. In other words, the estimates converge to the truth in probability. The notation $X_n \xrightarrow{p} Y$ for convergence in probability denotes $P(|X_n - Y| < \epsilon) \to 1$ as $n \to \infty$ for any $\epsilon > 0$. This allows us to state our consistency theorem. A consistency proof that is typical for $c\ell$ estimators is given in the appendix.

Theorem 1 (Consistency). *Given the standard asymptotic regime, $\hat{N}_{c\ell}/d \xrightarrow{p} \lambda$ is a consistent estimator of λ.*

The theory of M-estimators and composite likelihood also gives that $c\ell$ estimators have an asymptotically normal limit distribution given by

$$\mathcal{G}(n)^{1/2}(\hat{N}_{c\ell} - n) \rightsquigarrow N(0, 1) \tag{13}$$

where $\mathcal{G}(n) = \mathbb{E}\, c\ell''(n; C)^2 / \mathrm{Var}\, c\ell'(n, C)$ is called the Godambe information at n [22]. The notation $Y_n \rightsquigarrow N(0, 1)$

denotes convergence in distribution. Equivalently, the sequence of cdfs for Y_n converges to the cdf on the right. This distributional result can be interpreted as the cardinality estimate $\hat{N}_{c\ell}$ is approximately unbiased and normally distributed with $\operatorname{Var} \hat{N}_{c\ell} \approx \mathcal{G}(n)^{-1}$. Given this general result, the problem of obtaining an error estimate is reduced to the exercise of computing the necessary quantities. The primary challenge in computing them is that $c\ell'(n, C)$ is the sum of *dependent* random variables. Consider the variance term in the Godambe information.

$$
\begin{aligned}
s_n(C_b) &= \frac{\partial}{\partial n} \log f(C_b|n) = \frac{\partial f(C_b|n)/\partial n}{f(C_b|n)} \\
\operatorname{Var} c\ell'(n; C) &= \operatorname{Var}\left(\sum_b s_n(C_b)\right) \\
&= d\,\mathbb{E} s_n(C_1)^2 + d^2(d-1)\operatorname{Cov}(s_n(C_1), s_n(C_2))
\end{aligned} \tag{14}
$$

The first term of the last line is the variance when the registers are independent and is an easily computed upper bound. The second term captures the differences from independence. Unlike the independent component, one cannot simply substitute derivatives at the observed sketch values to compute this. That would lead to a degenerate estimate $\widehat{\operatorname{Var}}\, c\ell'(n; C) = 0$ since $c\ell'(\hat{n}, C)$ is always 0 at the maximizer $\hat{n}$.

Computing the covariance requires computing the full bivariate marginal distribution. Consider the bivariate cdf $\tilde{G}_2(x, y) := P(\tilde{Z}_1 \le x, \tilde{Z}_2 \le y)$ where $\tilde{Z}_i$ is defined in equation 7. Analogous to the derivation of the univariate maximum of random variables, the cdf and pmf of two registers is

$$
\begin{aligned}
F_2(x, y|n) &:= P(C_1 \le x, C_2 \le y) = \tilde{G}_2(x, y)^n \\
&= \left(1 - \frac{1}{d\,2^x} - \frac{1}{d\,2^y}\right)^n \\
f_2(x, y|n) &= F_2(x, y) - F_2(x-1, y) - F_2(x, y-1) + F_2(x-1, y-1)
\end{aligned}
$$

The covariance can be directly computed from these underlying probabilities. Algorithm 4 provides the complete procedure for a slightly more general case needed by our many distinct count sketches. Note that the Godambe error estimate is an a priori estimate, it can be computed ahead of time to speed computation or to establish error bounds for a given sketch size.

4 SKETCH CONSTRUCTION

We now show how to combine HLL sketches for distinct counting and Count-Min sketches that enable many additive counters to be stored in a small amount of space. We call the resulting family of sketches the Count-HLL sketches. In particular, we propose two new sketches, the pointwise Count-HLL sketch and the aggregate Count-HLL sketch. The former only allows for pointwise queries but has the strongest theoretical guarantees. The latter also allows for both types of aggregation queries that form the many distinct count problem.

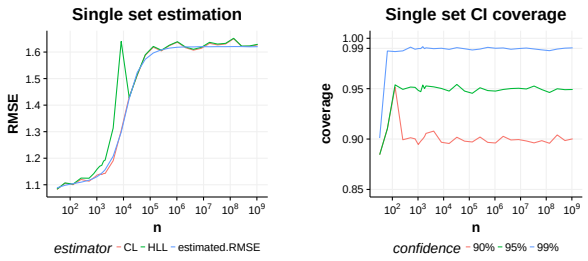

Figure 1: Cardinality estimation on a HLL sketch with 4096 bins. Left: Our estimator removes HLL's bias and our RMSE estimate matches the truth. Right: The confidence intervals deliver the promised coverage except for very small cardinalities where undercoverage is due to discreteness.

A critical idea for designing and making use of our sketches is a decomposition of the sketch register values into a standard HLL sketch register and an independent noise term that can be estimated from the sketch itself. This both enables application of the composite likelihood estimator and reasoning about what constitutes a good sketch design. Our sketches can be seen as a modification of Count-Min but takes a non-linear projection and has non-additive errors arising from using the maximum operation.

4.1 Pointwise query construction

Our simplest method uses independent hashes for all the labels. The main downside of this approach is that full independence makes further aggregations over labels impossible. For this construction, choose independent hash functions h, h', h'' such that for each pair (x_t, ℓ_t) in the stream

$$(x_t, \ell_t) \xrightarrow{h} R_t \sim Uniform(d) \tag{15}$$

$$(x_t, \ell_t) \xrightarrow{h'} V_t \sim Geometric(1/2) \tag{16}$$

$$(R_t, \ell_t) \xrightarrow{h''} B_t \sim Uniform(w) \tag{17}$$

and $\{R_t\}_t, \{B_t\}_t, \{V_t\}_t$ are all mutually independent. The sketch then takes the maximum value that each hashed to each entry. Formally,

$$C_{rb} := \max\{V_t : R_t = r, B_t = b\} \cup \{0\}. \tag{18}$$

This construction and others are expressed in pseudocode in algorithm 1.

4.2 Pointwise Background Distribution

We now show that given a label ℓ, the sketch registers for ℓ can be decomposed into a HLL sketch register and an independent error term. This error distribution Φ is a function of all the cardinalities other than n_ℓ. It cannot be described analytically without including them as parameters, and hence one cannot apply composite-likelihood estimation directly as there would be more parameters than observations. However, using the idea of a background distribution introduced by [31] for the CountMin sketch, we can show that this error distribution can be empirically estimated from the sketch itself. This yields a marginal likelihood that is only a function of the cardinality of interest and allow us to apply our machinery for estimation.

Denote by $C(\mathcal{D})$ the sketch constructed on the data set $\mathcal{D}$ of item, label pairs. For sketches C and C', denote by $C \wedge C'$ the entrywise maximum of C and C'. It is easy to see that one may decompose the sketch as

$$C(\mathcal{D}) = C(\mathcal{D}_\ell) \wedge C(\mathcal{D}_{-\ell}) \tag{19}$$

where $\mathcal{D}_\ell, \mathcal{D}_{-\ell}$ are subsets of the data $\mathcal{D}$ that include only elements with label ℓ or only those without label ℓ respectively. This immediately leads to the independence of the error term as the variables used to construct $C(\mathcal{D}_\ell)$ are independent from those used to construct $C(\mathcal{D}_{-\ell})$.

Lemma 2 (Independence of noise). *Under the pointwise query construction, the signal $C(\mathcal{D}_\ell)$ is independent from the noise $C(\mathcal{D}_{-\ell})$.*

Furthermore, by symmetry, all entries of $C(\mathcal{D}_{-\ell})$ are identically distributed. We write their common distribution as Φ and refer to it as the background distribution as it contains no information about the cardinality of interest.. Thus, the distribution of any register C_{rb} is given by

$$\epsilon_{rb} \sim \Phi \tag{20}$$

$$C_{rb} = C_{rb}(\mathcal{D}_\ell) \wedge \epsilon_{rb} \tag{21}$$

where Φ is the unknown background distribution and $C_{rb}(\mathcal{D}_\ell) \perp \perp \epsilon_{rb}$.

We estimate the background distribution Φ by simply tabulating all $d(w-1)$ values in the sketch that do not contain signal. Since there are at most 64 possible values for a 6 bit register, this is trivial to do and results in a consistent estimate of Φ. We will refer to this estimator as the *raw empirical estimator*.

Using the distribution of the maximum of random variables given in equation 4, the resulting cdf for C_{rb} is the product of the HLL cdf and background cdf:

$$\tilde{F}(v|n_\ell) = \tilde{G}(v)^{n_\ell}\Phi(v). \tag{22}$$

Substituting the corresponding pmf into the composite marginal log-likelihood in equation 11 and maximizing it gives the $c\ell$ estimator for this sketch. The proof of theorem 1 is easily generalized since a uniform convergence of the background distribution estimate preserves the uniform convergence of the composite-likelihood in a compact neighborhood of the truth. This gives the following theorem.

Theorem 3 (Pointwise construction consistency). *Suppose for all ℓ, $n_\ell/d \to \lambda_\ell$ as $d \to \infty$. Then the $c\ell$ estimator for the pointwise construction is a consistent estimator of λ_ℓ.*

4.3 Aggregation construction

In the pointwise construction, data points $(x, \ell), (x, \ell')$ sharing the same item x can be hashed to two different bins and have different values. This makes the sketch duplicate sensitive if an aggregation merges those two bins or if those bins correspond to separate registers in the merged sketch.

The aggregation construction removes these duplicate sensitive updates by making the row R_t and value V_t only a function of the item x_t and not the label ℓ_t. The process for choosing the bin within a row is identical to the pointwise construction. The marginal distributions of R_t, B_t, and V_t are also remain unchanged. The only difference is that the random variables are no longer mutually independent.

$$x_t \xrightarrow{h} (R_t, V_t). \tag{23}$$

$$(R_t, \ell_t) \xrightarrow{h'} B_t \tag{24}$$

Since the value V_t does not depend on the label, an item cannot modify the same register twice. Since an item will always hash to the same row and aggregations merge only bins within the same row, two labels cannot cause two separate registers to be modified in a aggregated sketch.

Algorithm 1 Sketch-update(C, x, ℓ)

$$digest \leftarrow \begin{cases} (x,\ell) & \text{if pointwise} \\ x & \text{otherwise} \end{cases}$$
$$z \leftarrow \textsc{TrailingZeros}(\textsc{hash}(digest, seed_2))$$
$$r_0 \leftarrow \textsc{hash}(digest, seed_1) \bmod d$$
$$r \leftarrow \begin{cases} \textsc{hash}(digest, seed_1) \bmod d & \text{if vHLL} \\ r_0 & \text{otherwise} \end{cases}$$
$$b \leftarrow \textsc{hash}((r,\ell), seed_3) \bmod w$$
$$C[r,b] \leftarrow \textsc{max}(C[r,b], z)$$

This construction can also reduce the noise in the sketch. Unlike the pointwise construction which has as many random tuples as there are distinct $(item, label)$ pairs, the aggregation construction only introduces as many random tuples as there are distinct items. Thus, the background distribution contains smaller values. Figure 2 shows this for the high overlap case described in the experiments.

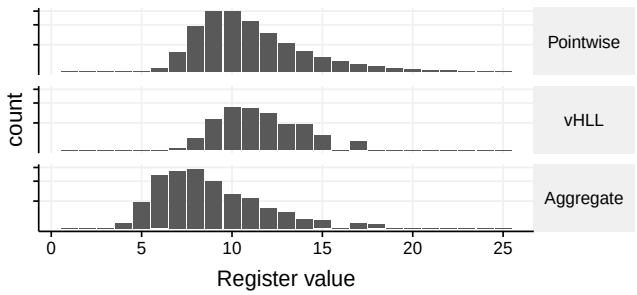

Figure 2: The aggregate construction has the background distribution with the smallest magnitudes.

However, estimation under this construction is more challenging as the background distribution is not directly observed. To see this, further split $\mathcal{D}_{-\ell}$, containing pairs with label not equal to ℓ, into $\mathcal{D}^+_{-\ell} := \{(x, \ell') \in \mathcal{D}_{-\ell} : x \in \mathcal{X}_\ell\}$ containing items that also appeared with label ℓ and its complement $\mathcal{D}^-_{-\ell}$. This yields a decomposition

$$C(\mathcal{D}) = C(\mathcal{D}_\ell) \wedge C(\mathcal{D}^+_{-\ell}) \wedge C(\mathcal{D}^-_{-\ell}). \tag{25}$$

Here, $C(\mathcal{D}^-_{-\ell}) \perp\!\!\!\perp C(\mathcal{D}_\ell)$ but $C(\mathcal{D}_\ell) \not\perp\!\!\!\perp C(\mathcal{D}^+_{-\ell})$ where $X \perp\!\!\!\perp Y$ denotes that X is independent of Y. In other words, the background distribution consists of two parts, one that is independent of the signal, and a dependent part induced by items shared with $\mathcal{D}_\ell$.

4.4 Aggregate Background Distribution

We address the problem of estimating the background distribution in three ways. First, we show that we can asymptotically recover the background distribution. However, the estimator does not have good finite sample properties. Second, we derive some simple bounds on the noise distribution. This allows us to provide bounds with theoretical guarantees. Finally, we heuristically derive an estimator with good finite sample performance.

4.4.1 Asymptotic consistency. We can estimate the background distribution consistently by identifying a set of rows that can be identified as draws from the noise distribution. This exploits the fact that, given a signal register, any register in the same row with value greater than it must contain noise. More precisely, fix a label ℓ. For any row r, let b_r be the column that ℓ hashes to in row r. Since the maximum hash value of items in $\mathcal{D}_\ell$ in row r is $\leq C_{rb_r}$, it follows that every entry in row r that is $> C(\mathcal{D})_{rb_r}$ must be equal to the entry from the background $C_{rb_r}(\mathcal{D}^-_{-\ell})$. This yields the following theorem.

Theorem 4 (Aggregation construction consistency). *Assume the width increasing asymptotic regime. Let $v_{min} = \min_r C_{rb_r}$ be the minimum value among the signal registers and $\mathcal{R} = \{r : C_{rb_r} = v_{min}\}$ be the collection of rows equal to that minimum. The empirical distribution of values in those rows $\mathbb{P}_{\mathcal{R}}(v) = \frac{1}{w|\mathcal{R}|}\sum_{r\in\mathcal{R}}\sum_{b'=1}^{w} 1(C_{rb'} \leq v)$ converges to the background distribution cdf, the marginal distribution of $C(\mathcal{D}^-_{-\ell})$.*

Algorithm 2 Estimate background distribution

function Ecdf(C)
  $T \leftarrow$ Tabulate(C)
  return CumulativeSum(T)/sum(T)
end function
function Estimate-Φ_{raw}($C^{(signal)}, C^{(noise)}, k_{label}$)
  $\hat{\Phi} \leftarrow$ Ecdf(C_{noise})
  return $x : x \to \hat{\Phi}_x^{k_{label}}$
end function
function Estimate-Φ_{agg}($C^{(signal)}, C^{(noise)}, k_{label}$)
  for $r = 1 \to d$ **do**
    $\mathbb{K}_r \leftarrow$ Ecdf($C^{(noise)}[r, :]$)
    $F_X^{(r)} \leftarrow \mathbb{K}_r^{k_{label}}$
    $F_{XX}^{(r)} \leftarrow \mathbb{K}_r^{2k_{label}}$
    $F_{XXY}^{(r)}[i] \leftarrow F_X^{(r)} 1(i < C[\mathcal{I}_r]) \quad \forall i = 1, \ldots, w$
  end for
  $\overline{F}_{XXY} \leftarrow \mathbb{E}_R F_{XXY}^{(R)}$ (average over rows R)
  $\overline{F}_{XX} \leftarrow \mathbb{E}_R F_{XX}^{(R)}$
  $\overline{F}_X \leftarrow \mathbb{E}_R F_X^{(R)}$
  $\overline{F}_{XY} \leftarrow$ Ecdf($C^{(signal)}$)
  $\hat{\Phi} \leftarrow \frac{\overline{F}_X^2 \overline{F}_{XXY}}{\overline{F}_{XX}\overline{F}_{XY}}$
  return $x : x \to \hat{\Phi}_x$
end function

Proof. Suppose there is a row r containing no items with label ℓ. Then, trivially the row $C_r(\mathcal{D}) = C_r(\mathcal{D}^-_{-\ell})$. Furthermore, since $C(\mathcal{D}_\ell) \perp\!\!\!\perp C(\mathcal{D}^-_{-\ell})$, conditioning on $C_r(\mathcal{D}_\ell) = 0$ does not affect the distribution of $C(\mathcal{D}^-_{-\ell})$. If $C_{rb} = 0$ then we have identified a row where there are no items from $\mathcal{D}_\ell$. Furthermore, the unconditional distribution of labels into bins in a row is given by $\tilde{M}^r \sim Multinomial(|\mathcal{L}|, 1/w, \ldots, 1/w)$. Conditional on $C_{rb_r} = 0$, the labels in $\tilde{M}^r_{b_r}$ that have items in row r are simply reallocated to the other bins. Since $|\mathcal{L}|/w \to \rho$, the proportion of affected bins goes to 0 in probability. Thus, the empirical distribution of the row conditional on $C_{rb_r} = 0$ converges in probability to the unconditional distribution. This unconditional distribution $\tilde{\Phi}$ satisfies $\mathbb{E}\tilde{\Phi} = \Phi$ by the definition of Φ. Since $P(C_{rb} = 0) \to \alpha > 0$ in the asymptotic regime, it follows that as $c_{min} = 0$ eventually and $d \to \infty$, $|\mathcal{R}| \to \infty$ as well. By the law of large numbers, the empirical distribution of the collection of rows $\mathcal{R}$ converges uniformly in probability to the limit background distribution. □

4.4.2 Bounds. Although theorem 4 asymptotically recovers the background distribution, it relies on an extremely low probability event, namely a signal register having value 0. Thus, the finite sample performance of the estimate may be poor. To obtain theoretical guarantees that hold in finite sample settings, we derive estimates of the noise distribution which are guaranteed to be either biased upwards everywhere or downwards everywhere. This can be used to derive practical bounds on the cardinality.

Since $C(\mathcal{D}_{-\ell}) \geq C(\mathcal{D}^{-}_{-\ell})$, the sketch entries contain more noise than the true background distribution. The *raw empirical estimator* applied to the aggregate construction thus overestimates the noise and leads to the following lemma.

Lemma 5. *Let Φ, Φ_0 be the marginal distributions of $C(\mathcal{D}^{-}_{-\ell})$ and $C(\mathcal{D}_{-\ell})$ respectively. Then $\Phi \geq \Phi_0$.*

When the noise estimate is too large, the cardinality estimates will correspondingly tend to be too low. By applying the Godambde variance estimate, one obtains an estimated lower bound for the cardinality.

Likewise, we find an estimate an upper bound on the cardinality by finding an upper bound on the noise distribution cdf. To do this, we counter-intuitively add even more noise to the signal registers and extrapolate the noise from the resulting change in distribution. We call the resulting estimator the SIMEX estimator due to its similarity to the Simulation-Extrapolation (SIMEX) method [10]. The upper bound is given by the following lemma.

Lemma 6. *Let $\tilde{b} \neq b_r$ be some bin. Let Q be the cdf of $C_{rb}(\mathcal{D}_{-\ell}) \wedge C_{r\tilde{b}}(\mathcal{D}_{-\ell})$, in other words the noise term if one added additional noise from another register on the same row. Then $\Phi \leq \frac{Q}{\Phi} = \frac{\hat{G}^{n_\ell} Q}{\hat{G}^{n_\ell} \Phi}$. In other words, one obtains a lower bound by taking the ratio of the distribution of the signal registers after adding more noise with the distribution of the signal registers.*

We prove this by examining the sets that are hashed to row r. Denote by Y the items with label ℓ that are hashed to row r, and let X, X' be the items from $\mathcal{D}_{-\ell}$ hashed to bins $b, \tilde{b}$ in row r. The noise distribution is given by $\Phi = \mathbb{E}G^{|X \backslash Y|}$ where G is the $Geometric(1/2)$ cdf. The noise distribution of the merged registers $Q = \mathbb{E}G^{|(X \cup Y) \backslash Z|}$. Since $|(X \cup X')\ backslash Y| \leq |X \backslash Y| + |X' \backslash Y|$, it follows that $\Phi^2 \leq Q$.

4.5 Bias corrected background estimator

The aggregation construction produces less noise than the pointwise construction. Despite this, estimators for the aggregation construction that are based on the bounds or asymptotic consistency result empirically perform worse as they have significant bias. We now give an estimator that allows the aggregation construction to beat the pointwise construction empirically.

Similar to the analysis for the SIMEX estimator, we examine the distribution after adding more noise. Let X, X', Y be defined as above in section 4.4.2. The inclusion-exclusion principle gives

$$\begin{aligned} |X' \cup X \cup Y| &= |X'| + |X| + |Y| - |X' \cap Y| - |X \cap Y| \\ &\quad - |X' \cap X| + |X' \cap X \cap Y| \\ |X' \cup Y| &= |X'| + |Y| - |X' \cap Y| \\ |X' \cup X| &= |X'| + |X| - |X' \cap X| \end{aligned}$$

The latter two equations can be used to cancel out the unwanted pairwise intersection in the first equation to obtain

$$\begin{aligned} &|X' \cup X \cup Y| - |X' \cup X| - |X' \cup Y| + |X'| + |X| \\ &= |X \backslash Y| - |X' \cap X \cap Y| \end{aligned}$$

To convert these to observable distributional quantities, we approximate

$$\Phi = \mathbb{E}G^{|X \backslash Y|} = \mathbb{E}G^{|X' \cup X \cup Y| - |X' \cup X| - |X' \cup Y| + |X'| + |X|} \tag{26}$$

$$\approx \frac{\mathbb{F}_{X \cup X' \cup Y} \mathbb{F}^2_X}{\mathbb{F}_{X \cup X'} \mathbb{F}_{X \cup Y}} \mathbb{E}G^{|X' \cap X \cap Y|} \tag{27}$$

where $\mathbb{F}_X$ is the raw empirical estimate of the error, $\mathbb{F}_{X \cup Y}$ is the empirical distribution of the signal registers, $\mathbb{F}_{X \cup X'}$ is that of two randomly merged non-signal registers in the same row, and $\mathbb{F}_{X' \cup X \cup Y}$ is that of the signal register merged with a random non-signal register in the same row. All of these are observable quantities as they are based on either a set of observed registers or on two merged registers. The approximation error arises from pulling the expectation into the product of non-independent random variables and using the empirical, observed quantities. To generate the estimator, we drop the last, unobserved quantity $\mathbb{E}G^{|X' \cap X \cap Y|}$ which represents the unremoved bias in the estimator. Dropping this term results in smaller noise estimates and an approximate downward bias of $\mathbb{E}|Y \cap X \cap X'|$ in the resulting estimated cardinalities In comparison, the raw empirical estimator has approximate bias $|X| - |X \backslash Y| = |X \cap Y|$.

4.5.1 Computation. Although $\mathbb{F}_{X \cup X'}$ is described as the empirical distribution obtained by randomly merging two registers, it is not necessary to perform the random merges to compute the distribution. Let $\mathbb{K}_r$ be the empirical cdf of the r^{th} row minus the signal register. For two random registers in row r, the empirical cdf of their maximum is $\mathbb{K}^2_r$. When the sketch width $w \to \infty$, the distributions for two random register drawn with and without replacement converge, so it may be replaced by the empirical cdf of the row when w is large. Thus, we can compute $\mathbb{F}_{X \cup X'} \xrightarrow{p} d^{-1} \sum_r \mathbb{K}^2_r$ under the uniform norm. Likewise, $\mathbb{F}_{X' \cup X \cup Y}(u) \xrightarrow{p} d^{-1} \sum_r 1(C_{rb_r} \leq u) \mathbb{K}_r(u)$ uses the empirical cdf $1(C_{rb_r} \leq u)$ of the solitary signal register in row r when computing the distribution of the maximum. This is summarized in algorithm 2.

4.5.2 Impossibility result. Disentangling the dependent and independent components of the background is not practically possible in all scenarios. Consider the case where there are many labels $|\mathcal{L}| \gg w \log d$ with each consisting of the same set of items $\mathcal{X}_{freq}$. In this case, for each row, all nonzero noise entries are equal to to the same value, and with high probability any collection of d noise entries are non-zero with high probability. If another set $\mathcal{X}_\ell$ is added to the sketch, $\mathcal{X}_\ell$ or $\mathcal{X}_{freq} \cup \mathcal{X}_\ell$ are practically impossible to distinguish. This is consistent with our inability to remove the three-way intersection term in the bias corrected estimate of the noise.

4.6 Aggregations and Mergeability

Three types of aggregation may be performed on the sketches. First, given two data streams $\mathcal{D}, \mathcal{D}'$, their corresponding sketches can trivially be merged

$$C(\mathcal{D} \cup \mathcal{D}') = C(\mathcal{D}) \wedge C(\mathcal{D}'). \tag{28}$$

This is especially useful in distributed settings or when computing aggregations over time. Sketches for many small time windows can be merged to generate a sketch for a longer window.

The other two aggregation types form the many distinct count problem and merge labels in a single sketch. The query for the total number of distinct items in a sketch is simple for the aggregation construction and impossible for the pointwise and vHLL constructions. The row-wise maximum of the aggregate construction is the same as a regular HLL sketch on all the items. For queries that merge a list of labels $\mathcal{J}$, the aggregate construction can merge them as long as there are enough remaining registers containing no signal in order to estimate the background distribution.

Given two labels ℓ, ℓ' with distinct item sets $\mathcal{X}_\ell, \mathcal{X}'_\ell$, denote the registers containing signal for ℓ as $C_\ell = \mathcal{S}(\mathcal{X}_\ell) \wedge \epsilon_\ell$ where $\mathcal{S}(\mathcal{X}_\ell)$ is the HLL sketch containing just items in $\mathcal{X}_\ell$. Likewise define $C_{\ell'}$ for ℓ'. Trivially

$$C_\ell \wedge C_{\ell'} = \mathcal{S}(\mathcal{X}_\ell \cup \mathcal{X}_{\ell'}) \wedge (\epsilon_\ell \wedge \epsilon_{\ell'}) \tag{29}$$

The error cdf for $\epsilon_\ell \wedge \epsilon_{\ell'}$ can be computed similarly to the procedure in section 4.5.1. The only differences are that the per row empirical cdf $\mathbb{K}_r$ is replaced by $\mathbb{K}_r^2$ and the signal register C_{rb} is replaced by the merged value $C_{rb} \wedge C_{rb'}$ where b, b' are the bins with signal in row r for ℓ and ℓ'.

5 ESTIMATION ALGORITHMS

Here we provide explicit algorithms for estimation of cardinalities, errors, and the background distribution. For cardinality estimation, we use a Newton-Raphson procedure for composite likelihood maximization. Lemma 7 derives the gradient and Hessian for the composite log-likelihood and shows it is log-concave so that optimization is fast and the solution is unique. The proof is straightforward differentiation and is given in the appendix. Algorithm 3 gives the procedure to estimate a count given the background distribution, and algorithm 5 gives the complete procedure given the sketch and set of labels being queried. From this is it easy to see that query results are generated by four distinct and conceptually simple steps.

Lemma 7. *The composite marginal log-likelihood is strictly concave and has first and second derivatives given by*

$$c\ell'(n; X) = \sum_i \log \tilde{G}(X_i) - \frac{\log r(X_i)\phi(X_i)}{(r(X_i)^{-n} - \phi(X_i))} \tag{30}$$

$$c\ell''(n; X) = \sum_i \frac{r(X_i)^{-n} \log^2 r(X_i)\phi(X_i)}{(r(X_i)^{-n} - \phi(X_i))^2} \tag{31}$$

where $r(v) = \tilde{G}(v-1)/\tilde{G}(v)$ *and* $\phi(v) = \Phi(v-1)/\Phi(v)$.

Algorithm 3 Max-$c\ell$-Estimator(S: signal registers, Φ)

$d \leftarrow length(S)$
for $x = 0 \rightarrow 64$ **do**
  $\tilde{G}_x \leftarrow 1 - 2^{-x}/d$
  $r_x \leftarrow \tilde{G}_{x-1}/\tilde{G}_x$
  $\phi_x \leftarrow \Phi(x-1)/\Phi(x)$
end for
$\mathbf{W} \leftarrow$ Tabulate(S)
$\hat{n} \leftarrow$ InitialGuess(S)
repeat
  $c\ell'(\hat{n}) \leftarrow \sum_{x=0}^{64} W[x] \left(\log \tilde{G}_x - \frac{\log r_x \phi_x}{r_x^{-\hat{n}} - \phi_x} \right)$
  $c\ell''(\hat{n}) \leftarrow \sum_{x=0}^{64} W[x] \frac{r_x^{-n} \log^2 r_x \phi_x}{(r_x^{-n} - \phi_x)^2}$
  $\hat{n} = \hat{n} - \frac{c\ell'(\hat{n})}{c\ell''(\hat{n})}$
until Convergence
return $\hat{n}$

5.1 Error estimates

Similar to estimation, the same error estimator for HLL can be used for Count-HLL except for a small change to the cdf. In this case, the bivariate cdf $\tilde{G}_2^{n_\ell}$ is replaced by its product with the error cdf $\tilde{G}_2^{n_\ell} \Phi_2$. The following lemma shows that the bivariate error cdf Φ_2 can be replaced by the product of the univariate error cdfs.

Lemma 8. *In the width increasing asymptotic regime, the bivariate error distribution for a label ℓ converges to the product of the marginal error distributions* $\Phi_2(x, y) \rightarrow \Phi(x)\Phi(y)$.

Proof. Let b_1, b_2 be two random bins in rows 1 and 2. The number of labels in each bin converge to independent $Poisson(\rho)$ random variables by the Poisson limit theorem. The probability that any of the labels in entry $(1, b_1)$ also

hashes to $(2, b_2)$ is $(L/w)L^{-1} = 1/w \to 0$. Therefore, the process converges to one where one independently draws $M_i \sim Poisson(\beta)$ labels for each bin and then independently draws M_i sets for each. Therefore, $\Phi_2(x, y) \to \Phi(x)\Phi(y)$ □

Figure 6 shows that the theoretically correct CI's deliver almost exactly the right coverage on real datasets whenever the estimates have RRMSE < 50%. For larger RRMSE, an exact CI is of little practical use as the interval is larger than the estimate itself. In difficult synthetic datasets, figure 7 shows the estimated standard deviations are accurate even when the estimates are biased and lead to CI's with undercoverage.

Algorithm 4 GodambeVariance($\hat{n}, \Phi, d$)

for all $x = 0 \to 64$ and $y = 0 \to 64$ **do**
  $G_{xy} \leftarrow (1 - 2^{-x}/d - 2^{-y}/d)$
  $F_{xy} \leftarrow G_{xy}^{\hat{n}} \Phi(x)\Phi(y) 1(x \geq 0) 1(y \geq 0)$
  $dF_{xy} \leftarrow \hat{n} F_{xy} \log G_{xy}$
end for
for all $x = 0 \to 64$ and $y = 0 \to 64$ **do**
  $f_{xy} \leftarrow F_{xy} - F_{x-1,y} - F_{x,y-1} + F_{x-1,y-1}$
  $s_{xy} \leftarrow \left(dF_{xy} - dF_{x-1,y} - dF_{x,y-1} + dF_{x-1,y-1}\right) / f_{xy}$
end for
$E \leftarrow \sum_{x,y} f_{xy} s_{xy}$
$V \leftarrow \sum_{x,y} f_{xy} s_{xx}^2$
return E^2/V

Algorithm 5 Query(C: sketch, Q: labels)

$C^{(signal)} \leftarrow \bigwedge_{\ell \in Q} C[\textsc{GetIndices}(\ell)]$ (vectorwise)
$\Phi \leftarrow \textsc{Estimate-}\Phi(C^{(signal)}, C, |Q|)$
$\hat{n} \leftarrow \textsc{Max-}c\ell\textsc{-Estimator}(C^{(signal)}, \Phi)$
$\hat{\sigma} \leftarrow \textsc{GodambeVariance}(\hat{n}, \Phi, nrow(C))^{1/2}$
return $(\hat{n}, \hat{\sigma})$

5.2 Time complexity

The update cost for both our constructions is $O(1)$. The time to tabulate the signal registers for any label is $O(d)$. Given this table T, the number of cdf evaluations needed in each Newton iteration to compute the composite likelihood or any of its derivatives is $O(|T|)$. If the initial estimate of the cardinality is good, then the Newton method has quadratic convergence and takes $O(\log\log c\epsilon^{-1})$ iterations where ϵ is the tolerated error [6] and c is some constant depending on second derivative of the marginal likelihood. The cost of estimating the background distribution from scratch takes $O(dw)$. If the values in the sketch are tabulated and maintained with the sketch, then the raw empirical estimate of the background can be computed in $O(|M|)$ time where M is the maximum register value. Since the maximum register value is typically stored in 5 or 6 bits, $|M| \leq 2^6 = 64$ may be treated as a constant. For the other background estimators for the aggregate construction, one must maintain a table for each row. The time to compute the background distribution is $O(d|M|)$. These tables can be maintained in $O(1)$ time. Asymptotically, $|M| = O_p(\log\log n_{tot})$ where n_{tot} is the total number of unique hash values. Here $V = O_p(f(n))$ is the probabilistic analog of big-O notation and denotes $P(V/f(n) < c) \to 1$ for some constant c as $n \to \infty$.

6 LIMITATIONS OF EXISTING METHODS

Past approaches for the many distinct count problem have yielded methods that naively combined the two (CM-FM) [9] or have no correctness guarantees (vHLL) [35]. We prove both of their constructions and estimators are flawed. They can have pathological behavior, and even in realistic scenarios, the estimators can have large bias. The proofs are deferred to the appendix.

6.1 CountMin-Flajolet-Martin

The CountMin Flajolet-Martin (CM-FM) sketch is identical to the CountMin sketch with the exact integer counters replaced with approximate distinct counting counters. Since each counter consists of m registers, we write the parameters of the CM-FM sketch as $d \times m \times w$. An equivalently sized Count-HLL sketch has parameters $dm \times w$. The inefficiencies in the CM-FM sketch primarily arise from the use of the minimum to combine counters and the limited collision resistance offered by the sketch design. While CountMin ensures errors are one-sided errors so that taking the minimum can never increase the error, the CM-FM sketch can perversely perform worse as more memory is allocated to the sketch. In particular, it underestimates with estimates eventually approaching 0 as the sketch depth increases.

Theorem 9 (CM-FM bias). *Consider a fixed dataset $\mathcal{D}$. Let $\hat{n}_\ell^{(CMFM)}$ be the cardinality estimate for label ℓ using a CMFM sketch. If the sketch depth $d \to \infty$ while the width and the size of individual counters stays the same, then $p(\hat{n}_\ell \leq 2) \to 1$.*

For this reason, the theoretical guarantees for the CM-FM sketch assume that the depth is small. However, the problem still exists with shallow sketches. For example, in the simple case when there is no noise from collisions, the CM-FM estimator is strictly worse than a single HLL estimator.

Theorem 10 (CM-FM inefficiency). *In the noiseless setting, if $m \to \infty$ and the cardinality estimates have an asymptotically normal distribution centered on the truth, then the CM-FM estimate is never better than using a single sketch in expectation. That is $\lim_{m\to\infty} P((\min_j \hat{n}^{(j)} - n)^2 > \epsilon_m) > \lim_{m\to\infty} P((\hat{n}^{(i)} - n)^2 > \epsilon_m)$ for any sequence ϵ_m where the r.h.s. converges to some value $p > 0$.*

These results point out issues in the estimation algorithm. More fundamentally, the sketch design is inherently less collision resistant than our Count-HLL designs. A label can tolerate up to d collisions with the CM-FM sketch while Count-HLL can tolerate up to $d \cdot m$ collisions. Figure 3 shows the performance of the CM-FM sketch degrades quickly once collisions become frequent. Equal sized Count-HLL sketches, on the other hand, degrade gracefully.

6.2 Virtual HyperLogLog

The virtual HyperLogLog sketch (vHLL) construction [35] is similar to our aggregate construction but does not ensure that the same item with different labels is always hashed to the same row. This prevents it from being able to aggregate over all labels or many labels. Like CM-FM, the estimator for vHLL has poor properties. This estimator is derived under an incorrect assumption of additive errors and can result in grossly biased estimates for real data sets. When the estimator is replaced by a composite likelihood estimator, the construction becomes practically useful even though it still has several poor theoretical properties. Empirically, we find the VHLL construction paired with the $c\ell$ estimator performs similarly to the pointwise construction and estimator but has the added ability to aggregate over a limited list of labels.

Rather than choosing the d sketch entries for each label row-wise, vHLL chooses them completely at random. This is equivalent to choosing a hash $h' : (original\ row, label) \to (new\ row, new\ bin)$ that adds another layer of randomization to the row when compared to the Count-HLL aggregate construction.

$$x_t \xrightarrow{h} (V_t, R'_t) \tag{32}$$

$$(R'_t, \ell_t) \xrightarrow{h'} (R_t, B_t) \tag{33}$$

Denote by $\mathcal{I}(\ell)$ the sketch indices that label ℓ hashes to. For estimation, the vHLL estimator is $\gamma(\hat{N}^{(HLL)}(C_{\mathcal{I}(\ell)})) - \hat{N}_{total}/w)$ where $\hat{N}^{(HLL)}$ is a single set HLL estimator, $\gamma = dw/(dw - d)$, and $\hat{N}_{total}$ estimates of cardinality of all items. They suggest using $\hat{N}_{total} = \hat{N}^{(HLL)}(C)$ where the entire sketch is treated as a single HLL sketch with dw entries. The formula treats the bins $C_{\mathcal{I}(\ell)}$ as an HLL sketch whose estimate should be debiased by the average distinct count per bin n_{total}/m. This formula uses an incorrect implicit assumption that adding k items to a single bin will increase the HLL cardinality estimate by k in expectation.

Theorem 11 proves this incorrect assumption can always lead to large bias regardless of the bias estimator. For the specific proposals for a bias estimate used in [35], the bias can be made arbitrarily large. These proofs themselves suggest that when there are heavy hitters, the vHLL estimator returns highly biased results unless the sketch is effectively empty; that is there are few sets with cardinality similar to or larger than the one being queried. This behavior is confirmed empirically in figures 3 on synthetic data and 5 on real PubMed data. It does not appear in the ad data.

Theorem 11 (Incorrect vHLL linearity assumption). *For any $\delta > 0$ and distinct count n_ℓ, there exists a size Z such that for any sketch with size $d \times w > Z$, the bias of any vHLL estimator $\mathbb{E}\hat{n}_\ell - n_\ell \geq n_\ell(1-\delta)$ for some data set $\mathcal{D}$ containing n_ℓ distinct items with label ℓ.*

Theorem 12 (vHLL bias and inconsistency). *The vHLL sketch can have arbitrarily large relative bias when applying the specific proposals to use the HLL estimate on the whole sketch or the true distinct count to debias the raw HLL estimate.*

The bias is a property of the estimator and can be fixed with a better estimator. However, there remain some fundamental issues due to the randomization procedure.

Theorem 13 (vHLL pathological variance). *For any unbiased estimator $\hat{n}$ on a vHLL sketch, $\hat{n}_\ell$ has infinite variance for some data set.*

The intuition behind this is that when a single item has a large hash value and is hashed to every register, then the sketch may contain only information about that item. This lack of information results in high variance. This may happen, for example, if there are many labels and every label's set contained `NULL`.

7 EXPERIMENTS

We run experiments on real and synthetic data to demonstrate (1) the state-of-the-art performance of our sketches and estimators, (2) the accuracy of our theory and error estimates even in finite sample regimes, and (3) practical considerations in implementing and sizing the sketches. We highlight that the asymptotic theory nearly perfectly predicts the performance of our sketches. Not only does it perfectly predict the bias or lack thereof of our sketches, but it also provides tight confidence intervals and perfectly predicts the variance of the estimates. It does so under all scenarios. Thus, practitioners can have confidence that the methods will generalize well to any scenario, and especially on larger datasets where the asymptotic theory is even more accurate.

Our primary metric for evaluation is the Relative Root Mean Squared Error (RRMSE) . For two sketches $\hat{C}, \tilde{C}$ with estimators $\hat{N}_\ell, \tilde{N}_\ell$, this is defined by

$$\text{RRMSE}(\hat{N}_\ell) = \sqrt{\mathbb{E}\left(\hat{N}_\ell/n_\ell - 1\right)^2} \times 100\% \tag{34}$$

We also consider the relative efficiency (RE), the ratio of relative MSEs. An estimator with RE 0.5 compared to a another requires twice the space to achieve the same error when both estimators are asymptotically normal and unbiased. Because of this we define the *relative size* to be the inverse of the RE.

We compare our sketch to vHLL [35] and CM-FM [9] and two variants vHLL* and CM-FM*. vHLL* uses the exact count of distinct items in the entire sketch rather than an estimate. CM-FM* uses d copies of a single HLL sketch to remove the bias from taking the minimum of independent estimates. For CM-FM, we use HLL counters instead of inefficient PCSA ones to make our results more comparable.

Our synthetic experiments cover two scenarios: when a sketch is "filled" by increasing the number of labels and when sets have high overlap. The first case is shown in figure 3. We simulate a datasets with n_{label} labels with n_{label} ranging from 100 to 3200. Each sketch contains 1024×1000 registers and is considered "saturated" when it has more labels than it has columns. The CM-FM sketches have depth $d = 4$. Each label contains 10^6 items. The item sets are all disjoint, so the pointwise and aggregate constructions are the same. We use the raw empirical background estimate.

Figure 3 shows Count-HLL dominates other methods, typically by over an order of magnitude. It is the only method with no detectable bias. It also has the smallest variance, no extreme estimates, and behavior that smoothly degrades. Both vHLL's and CM-FM's performance degrades badly as more sets are added. The poor performance of vHLL is due to bias and manifests even when the sketch is not even close to capacity. CM-FM degrades when there are collisions in all d bins. This only happens when the sketch is almost saturated. In the easy, low noise regime where the number of labels $|\mathcal{L}| = 100 \ll w = 1000$, only Count-HLL performs as well as individual HLL sketches. CM-FM requires roughly $d = 4$ times the space of Count-HLL as predicted by the theory. We will omit CM-FM from further comparisons since other experiments yield a high rate of hash collisions where CM-FM is guaranteed to perform poorly on.

In the second case, figure 4 examines the performance when there is high overlap among the sets with different labels. The data is drawn from a universe of 10^6 items. From these we fix a set of 10% as common items. We draw $k_{small} = 10^5$ or 10^6 sets of size 1000 from the common items and 1000 sets of size n_{big} from the full universe. We then estimate the cardinality for each of the large sets.

Our Count-HLL sketches always perform the best. vHLL typically has high bias. In the case where there are a moderate number $k_{small} = 10^5$ of background labels, the bias and relative error for vHLL degrades as the true cardinality increases, despite the signal being better separated from the noise. In contrast, our methods perform better as one would expect. The aggregate construction always has the lowest standard deviation but is biased downwards. The smaller standard deviation results in improved RRMSE when the variance dominates the bias, particularly in hard to estimate cases with smaller cardinalities. The case where $k_{small} = 10^6 \gg w$ mimics the challenging case where each bin contains almost all the common items for its row. The bias in this case asymptotes at 10%, precisely the proportion of common items and the amount of bias one would predict for the aggregate construction. When the true cardinalities are large, the pointwise construction is able to better the aggregate construction since the bias dominates the variance. The dashed lines in the rightmost figure show the raw empirical and the SIMEX estimators which are downwardly and upwardly biased as suggested by the theory. However, the bounds they produce can be quite loose.

We consider two real datasets, the PubMed bag of words dataset [13] and the KASANDR ad impression data set [27], representing realistic problems in natural language processing and advertiser reporting. For the PubMed dataset we count the distinct documents each word appears in for the top 8000 words. For PubMed, there are 141K distinct words and 8.2M documents with 730M word, document pairs. For the ad data set, we estimate the number of distinct users that have seen each of the 1000 most seen ads. The ad data contains 291K users and 2.2M ads with 15.8M pairs. We implemented sketch updates in C++ using the hash function Murmurhash3_x64_128. For simplicity, we used 8 bits per sketch entry. We were able to process 15K items/ms on an Intel Xeon E5-2430V3 2.4Ghz processor in a single thread. The estimation algorithms were implemented in R with optimization carried out by `nlm`.

Figure 5 shows the RRMSE on these real datasets for a variety of sketch sizes. Our Count-HLL sketches dominate vHLL. For the PubMed dataset, vHLL's accuracy does not significantly improve when increasing the sketch size by increasing the depth. This is since increasing the depth does not change the bias of the vHLL estimator. For the ad data, vHLL is worse, but competitive. In both cases, the aggregate and pointwise constructions perform extremely similarly with completely indistinguishable RRMSE curves except at narrow widths where the aggregate construction holds a slight edge. The plots also show the number of registers needed to achieve a given accuracy using individual HLL sketches and the actual number of registers per label. Our methods use $< 1\%$ and $< 10\%$ of the space to achieve the same error. When examining the effect of sketch parameters, each curve traces out the RRMSE as the depth and width are varied for a given space budget. The plots suggest a good rule of thumb is to choose the widest possible sketch that still satisfies $1/\sqrt{d} < c\,\epsilon$ where ϵ is the target RRMSE and $c < 1$ is some constant fudge factor. In all regimes, increasing the width does not hurt much unless it causes the depth to be fundamentally too small to reach the target error, but it can significantly help when the sketch is too narrow. We also found that when *vHLL* is paired with our $c\ell$ estimator

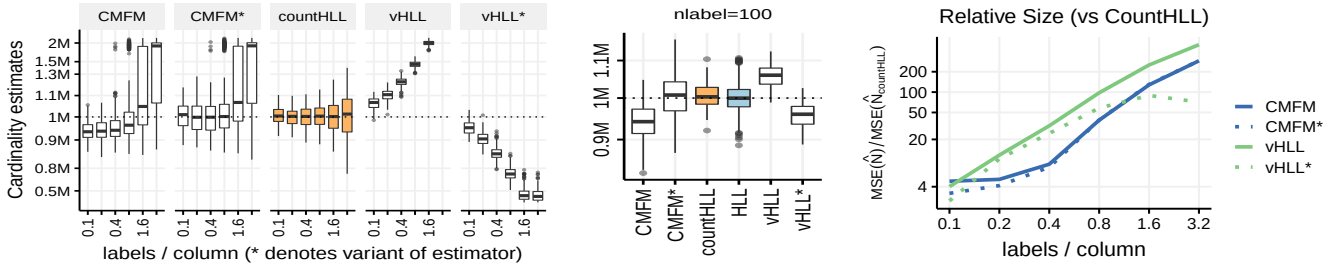

Figure 3: Effect of saturating a sketch.

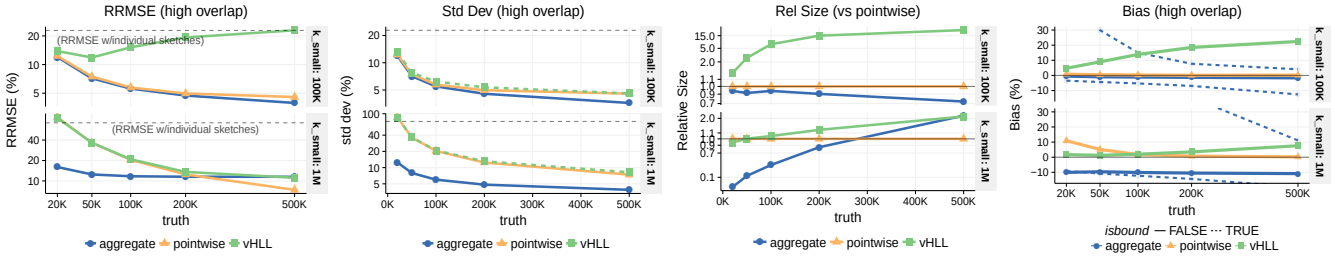

Figure 4: Comparison on sets with high overlap. The dashed gray lines in the left figures show the performance if using individual HLL sketches with the same space budget.

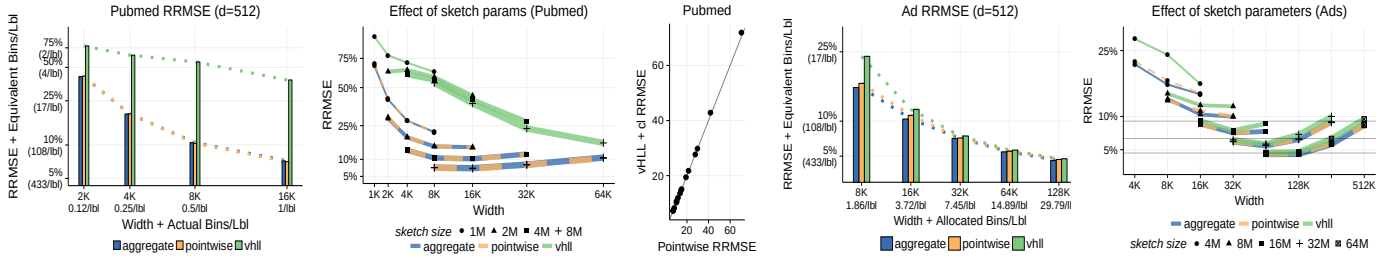

Figure 5: Results on the PubMed bag of words and KASANDAR ad datasets.

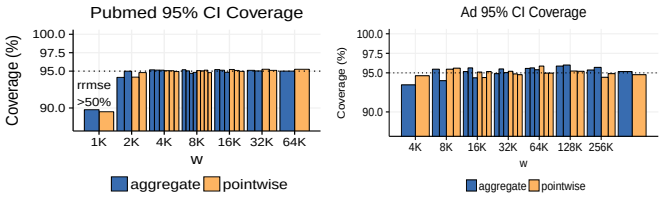

Figure 6: Confidence intervals are tight whenever estimates are useful

using the raw empirical background estimator, the RRMSE closely matches that of the pointwise construction.

8 DISCUSSION

We tackle approximate distinct counting for many datasets under a strict space budget while retaining many of the properties of the strategy of storing individual HLL sketches for each dataset. Much of our work focuses on correct estimation of the background distribution and cardinalities. This allows us to create the aggregate construction which performs well except in cases of extreme overlap. We note, however, that applying our estimation methods to the vHLL construction also yields good empirical results and may be a good option for those interested in only partially solving the many distinct count problem without theoretical guarantees. For future work, we believe the design of the aggregate construction can lead to other interesting designs, such as cases where there may be a natural hierarchy of aggregations. Furthermore, it may be possible to estimate an average bias for the aggregate construction and further improve the overall RRMSE. One last area of future work is addressing the question of asymptotic efficiency as it is the one property of the MLE that is not provably retained by our $c\ell$ estimator.

REFERENCES

[1] N. Alon, Y. Matias, and M. Szegedy. 1999. The Space Complexity of Approximating the Frequency Moments. *J. Comput. System Sci.* 58, 1 (1999), 137–147.

[2] Lars Backstrom, Paolo Boldi, Marco Rosa, Johan Ugander, and Sebastiano Vigna. 2012. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*. ACM, 33–42.

[3] Ran Ben Basat, Gil Einziger, Roy Friedman, and Yaron Kassner. 2017. Randomized admission policy for efficient top-k and frequency estimation. In *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*. IEEE, 1–9.

[4] James O Berger, Robert L Wolpert, MJ Bayarri, MH DeGroot, Bruce M Hill, David A Lane, and Lucien LeCam. 1988. The likelihood principle. *Lecture notes-Monograph series* 6 (1988), iii–199.

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

[6] Stephen Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

[7] E. Cohen. 2014. All-distances sketches, revisited: HIP estimators for massive graphs analysis. In *PODS*.

[8] Edith Cohen and Haim Kaplan. 2007. Summarizing data using bottom-k sketches. In *PODC*.

[9] Jeffrey Considine, Marios Hadjieleftheriou, Feifei Li, John Byers, and George Kollios. 2009. Robust approximate aggregation in sensor data management systems. *ACM Transactions on Database Systems (TODS)* 34, 1 (2009), 6.

[10] John R Cook and Leonard A Stefanski. 1994. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical association* 89, 428 (1994), 1314–1328.

[11] Graham Cormode and Minos Garofalakis. 2005. Sketching streams through the net: Distributed approximate query tracking. In *Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 13–24.

[12] G. Cormode and S. Muthukrishnan. 2005. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms* 55, 1 (2005), 58–75.

[13] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[14] Bradley Efron and David V Hinkley. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika* 65, 3 (1978), 457–483.

[15] Otmar Ertl. 2017. New cardinality estimation algorithms for HyperLogLog sketches. *arXiv preprint arXiv:1702.01284* (2017).

[16] C. Estan, G. Varghese, and M. Fisk. 2003. Bitmap algorithms for counting active flows on high speed links. In *Internet Measurement Conference*.

[17] P. Flajolet, É. Fusy, O. Gandouet, and F. Meunier. 2007. HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm. In *AofA*.

[18] Philippe Flajolet and G Nigel Martin. 1985. Probabilistic counting algorithms for data base applications. *Journal of computer and system sciences* 31, 2 (1985), 182–209.

[19] S. Heule, M. Nunkesser, and A. Hall. 2013. HyperLogLog in Practice: Algorithmic Engineering of a State of The Art Cardinality Estimation Algorithm. In *EDBT*.

[20] D. Kane, J. Nelson, and D. Woodruff. 2010. An optimal algorithm for the distinct elements problem. In *PODS*.

[21] Kevin J Lang. 2017. Back to the Future: an Even More Nearly Optimal Cardinality Estimation Algorithm. *arXiv preprint arXiv:1708.06839* (2017).

[22] B. Lindsay. 1988. Composite Likelihood Methods. *Contemp. Math.* (1988), 221–239.

[23] G. Manku and R. Motwani. 2002. Approximate frequency counts over data streams. In *VLDB*.

[24] A. Metwally, D. Agrawal, and A. E. Abbadi. 2008. Why go logarithmic if we can go linear?: Towards effective distinct counting of search traffic. In *EDBT*.

[25] A. Metwally, D. Agrawal, and A. El Abbadi. 2005. Efficient computation of frequent and top-k elements in data streams. In *ICDT*.

[26] J. Misra and D. Gries. 1982. Finding repeated elements. *Science of computer programming* 2, 2 (1982), 143–152.

[27] Sumit Sidana, Charlotte Laclau, Massih R Amini, Gilles Vandelle, and André Bois-Crettez. 2017. KASANDR: A Large-Scale Dataset with Implicit Feedback for Recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1245–1248.

[28] M. Thorup and Y. Zhang. 2004. Tabulation based 4-universal hashing with applications to second moment estimation.. In *SODA*, Vol. 4. 615–624.

[29] D. Ting. 2014. Streamed approximate counting of distinct elements: beating optimal batch methods. In *KDD*.

[30] Daniel Ting. 2016. Towards optimal cardinality estimation of unions and intersections with sketches. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1195–1204.

[31] D. Ting. 2018. Count-Min: Optimal Estimation and Tight Error Bounds using Empirical Error Distributions. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2319–2328.

[32] Daniel Ting. 2018. Data Sketches for Disaggregated Subset Sum and Frequent Item Estimation. In *Proceedings of the 2018 International Conference on Management of Data*. ACM, 1129–1140.

[33] Cristiano Varin. 2008. On composite marginal likelihoods. *Advances in Statistical Analysis* 92, 1 (2008), 1–28.

[34] K.Y. Whang, B. Vander-Zanden, and H. Taylor. 1990. A linear-time probabilistic counting algorithm for database applications. *TODS* (1990).

[35] Qingjun Xiao, Shigang Chen, Min Chen, and Yibei Ling. 2015. Hyper-Compact Virtual Estimators for Big Network Data Based on Register Sharing. *SIGMETRICS Perform. Eval. Rev.* 43, 1 (June 2015), 417–428. https://doi.org/10.1145/2796314.2745870

A TABLE OF SYMBOLS

h, h'	Hash functions
C	Matrix or array of sketch values
d, w	Sketch depth and width
(x, ℓ)	Item and label
n_ℓ	Cardinality of subset with label ℓ
$\hat{n}, \hat{n}_\ell$	Cardinality estimates
$\mathcal{L}$	Universe of labels
λ_ℓ	Asymptotic value of n_ℓ/d
ρ	Asymptotic number of labels per column $\lvert\mathcal{L}\rvert/w$
Φ	Background error distribution
$\mathbb{K}_r$	Empirical distribution for row r in C
$\mathcal{D}, \mathcal{D}_\ell, \mathcal{D}_{-\ell}$	Data, data with label ℓ, data without ℓ
$\mathcal{X}_\ell$	Set of unlabelled items in $\mathcal{D}_\ell$
$C(\mathcal{D})$	Sketch constructed from $\mathcal{D}$
$\mathcal{I}(\ell)$	Indices that items with label ℓ are hashed to
G	$Geometric(1/2)$ cdf
$\tilde{G}$	Marginal cdf for one item's value in fixed bin
R_t, B_t	Row and Bin (column) that (x_t, ℓ_t) hashes to
V_t	Hash value for (x_t, ℓ_t)
$c\ell$	Composite likelihood
h, h', h''	Hash functions
$C \wedge C'$	Entrywise maximum of sketches

Table 1: Table of symbols

B STATISTICAL CONCEPTS

B.1 Maximum likelihood estimation

One of the most common statistical parameter estimation techniques is maximum likelihood estimation. Under mild regularity conditions, it has a number of important properties including consistency and asymptotic efficiency. Given the true value of a parameter θ_0 and data $\mathcal{D}_t$, consistency is the property that the estimate converges the the truth in probability, $\hat{\theta}(\mathcal{D}_t)) \xrightarrow{p} \theta_0$, as the number of observations $|\mathcal{D}_t| \to \infty$. Asymptotic efficiency is the property that the estimator has the lowest asymptotic variance amongst *all* estimators. In other words, for any other estimator $\tilde{\theta}$ the limit $\lim_{t\to\infty} \operatorname{Var} \tilde{\theta}(\mathcal{D}_t)/\operatorname{Var} \hat{\theta}(\mathcal{D}_t) \geq 1$. Thus, not only does the maximum likelihood estimator achieve the best possible error rate, it achieves the best possible constant in front of that rate.

The probability of seeing the data for a parameter θ is called the likelihood $\mathcal{L}(\theta; \mathcal{D}) = p(\mathcal{D}|\theta)$. The maximum likelihood estimate (MLE) is simply the parameter that maximizes the probability of seeing the data.

$$\hat{\theta}_{MLE} = \arg\max_\theta \mathcal{L}(\theta; \mathcal{D}). \tag{35}$$

Due to both numerical issues and for the error estimation purposes described below, the equivalent log-likelihood maximization problem is solved instead. Here $\ell(\theta; \mathcal{D}) = \log \mathcal{L}(\theta; \mathcal{D})$ replaces the likelihood in the maximization and is equivalent since the log function is strictly increasing.

One of the main challenges with maximum likelihood estimation is ensuring that one can derive and compute the likelihood. We employ a generalization of maximum likelihood that enables easy computation.

B.2 Asymptotic Normality and Error estimates

In addition to achieving the best possible variance, the distribution of the MLE converges to a normal distribution with mean equal to the true value and variance given by the inverse Fisher information $I_t(\theta)^{-1}$. The Fisher information is $I_t(\theta) := \mathbb{E}\ell''(\theta; \mathcal{D}_t)$, and can often be replaced by the observed Fisher information $I_{obs} := \ell''(\hat{\theta}_{MLE}; \mathcal{D}_t)$ [14]. This results in asymptotically tight confidence intervals (CIs). Since $\hat{\theta}_t \approx Normal(\theta, I_t(\theta)^{-1})$, the interval $\hat{\theta}_t \pm 1.96 I_{obs}^{-1/2}$ yields an asymptotically tight 95% CI.

C PROOFS

Proof of theorem 1.

Proof. Denote $f_\theta = f(\cdot|n = \theta, m)$. Consider the Kullbeck-Liebler (KL) divergence $KL(f_n \| f_{\hat{n}})$. By the properties of the KL-divergence, $KL(f_n || f_{\hat{n}}) = 0$ if and only if $f_n = f_{\hat{n}}$. Furthermore, this only holds when $n = \hat{n}$. Since maximizing the expected log likelihood $J(\theta) = \mathbb{E} \log f(C_b|\theta, d)$ is equivalent to minimizing the KL divergence, it follows that the maximizer of J is n. To complete the proof, one must show that the empirical estimate $\frac{1}{d}\sum_{b=1}^{d} \log f(C_b|\theta, d)$ converges uniformly in probability to the true expectation for θ/d in a neighborhood of n/d. We omit the details, but this is easily accomplished replacing the underlying multinomial counts of distinct items per bin with a Poisson processes on d bins with rate n/d in each bin. The difference in the total count of this Poisson process and n is $O_p(\sqrt{n})$. Thus, at most $O_p(\sqrt{n}) = o_p(d)$ bins that can differ in value. Bounding this remainder term and noting that under the Poisson distribution, the terms are independent and the limit converges uniformly by Glivenko-Cantelli completes the proof. □

Proof of theorem 9

Proof. Each row in the CM-FM sketch is independent identically distributed by the same argument in [31] for the CountMin sketch. Let $C_{rj}^{(true)}$ be the true count of distinct items hashed to entry (i, j). If b_r is the bin in row r corresponding to ℓ, then $C_{rb_r}^{(true)} = n_\ell + \epsilon_r$ for i.i.d. errors ϵ_r. Let $n_{upper} > n_\ell$ be some value such that $P(C_{rb_r}^{(true)} <$

$n_{upper}) > 1/2$ and let $\hat{N}(n)$ be a random variable denoting an estimated cardinality when the true cardinality is n. Write $\delta := P(\hat{N}(C_{rb_r}^{(true)}) \leq 2) \geq P(\hat{N}(n_{upper}) \leq 2)/2 > 0$. Since the expected number of times d sketches return an estimate $\leq$ is $d\delta \to \infty$, it follows from the law of large numbers that $p(\hat{n}_\ell \leq 2) \to 1$. □

Lemma 14. *Let $\{\hat{n}^{(i)}\}_{i=1}^d$ be a collection of i.i.d. cardinality estimates on a set of n distinct elements. If the distribution $\hat{n}^{(i)} - n$ is symmetric about 0 and not degenerate, then for any i, $\mathsf{MSE}(\min_j \hat{n}^{(j)}) > \mathsf{MSE}(\hat{n}^{(i)})$ and $P((\min_j \hat{n}^{(j)} - n)^2 > \epsilon) > P((\hat{n}^{(i)} - n)^2 > \epsilon)$ for any ϵ where the probability on the right is non-zero.*

Proof. Let D_- be the number of estimates less than the truth n. Let $\delta_i = (\hat{n}^{(i)} - n)^2$ and $\delta_{(j)}$ be the j^{th} smallest value of the squared errors amongst the D_- estimates less than n. In other words, $\delta_{(j)}$ denotes the j^{th} order statistic. Let Ψ denote the cumulative distribution function of δ_i. By symmetry of the distribution, Ψ is also the conditional cdf given $\hat{n}_i \leq n$. By the inverse cdf transform, the distribution of $\delta_{(i)} \stackrel{d}{=} \Psi^{-1}(U_{(i)})$ where $\{U_{(i)}\}$ denote the order statistics for D_- uniform draws. Since Ψ^{-1} is monotone, the smallest estimate corresponding to U_{D_-} has the largest error, and hence larger error than the expectation $\mathbb{E}(\Psi^{-1}(U) - n)^2 = \mathbb{E}(\delta_i)$. The exact same argument holds with expectations replaced with probability statements. □

Proof of theorem 10

Proof. This follows immediately from lemma 14 in the appendix and symmetry of the normal distribution. □

Proof of theorem 11.

Proof. Consider the case where a stream consisting of $|\mathcal{L}| \geq w \log 2$ extremely large sets of size T plus one large set and one moderately large set of interest. [29] showed that the raw HLL estimator can be expressed simply as ρ/P_{update} where P_{update} is the probability a new distinct item updates the sketch. Following the same argument used for Bloom filters, the fraction of cells containing a large count concentrates on $1/2$. Taking T sufficiently large, the probability of a given cell being updated can be made arbitrarily small. Thus, the probability of update is essentially halved for any set of interest. Specifically, $P_{update} = P_{update}^{HLL}/2 + o_p(1)$ where P_{update}^{HLL} is the probability that a single set HLL estimator is updated.

This makes it impossible to provide good estimates for two different cardinalities. If $\hat{mu} = \hat{N}_{HLL}(C_\cdot)$ is the estimated bias term and C_{large}, C_{mid} denote collections of bins for a large set and the moderately sized set with sizes n_{large}, n_{mid} respectively, then one cannot solve

$$n_{large} = 2n_{large}(1 + o(1)) - \mathbb{E}\hat{n}_{bias} \tag{36}$$

$$n_{mid} = 2n_{mid}(1 + o(1)) - \mathbb{E}\hat{n}_{bias}. \tag{37}$$

To ensure the bias of $\hat{n}_{large}$ is less than n_{large}, one must take $\mathbb{E}\hat{n}_{bias} > n_{large}$. In this case, the absolute bias of $\hat{n}_{mid}$ is $> n_{large} - n_{mid}(1 + o(1))$ which is $> n_{mid}(1 + o(1))$ whenever $n_{large} > 2n_{mid}(1 + o(1))$. Therefore, one of the two must have relative bias exceeding $1 - \delta$. □

Proof of theorem 12.

Proof. For the specific proposal of using the true distinct count or the entire sketch to debias the estimate, we can precisely compute the bias. Suppose the sets are all disjoint. There are thus $\approx Tw \log 2$ distinct items and a true bias of $T \log 2$. This leads to estimates $\hat{n}_{large} \approx 2n_{large} - T \log 2 \approx 1.3T$ and $\hat{n}_{mid} \approx 2n_{mid} - T \log 2 < 0$ when the true total cardinality is used. One may hope that the problem may be mitigated when using the suggested estimator based off the entire sketch. We show that this does not help. When the sketch is wide, approximately half of the bins contain exactly 0, and the other half have nearly 0 probability of being updated. The raw HLL estimator will thus return an estimate of $\approx 1.5dw$. Since this falls into the small estimate regime for the HLL estimator, the estimate based on Linear Probabilistic Counting [34] will return $dw \log 2$. In both cases the □

Proof of theorem 13.

Proof. Consider the stream consisting of a single item x and $dw \log 2$ unique labels. Again, in expectation, half the entries in the sketch are filled. Furthermore, all entries contain the exact same value V. There is some non-zero probability δ for a given label ℓ, that all registers $C_{\mathcal{I}(\ell)}$ contain V. In this case, any estimator must be purely a function f of V. An unbiased estimator then satisfies $f(V)2^{-V}\delta + (1 - \delta)c = 1$ where c is the expectation of the estimator given not all registers contain V. Solving for $f(V)$ gives that $f(V) = \alpha 2^V$ for some constant $\alpha > 0$. Since $V \sim Geometric(1/2)$, the expectation is $\sum_{v=1}^{\infty} \alpha 2^v 2^{-v} = \infty$. Thus conditional on all registers for ℓ being non-zero, the estimator has infinite first moment as well as variance. By the law of total variance, the unconditional variance is infinite. □

Proof of theorem 7.

Proof. The log-marginal probability and its first and second derivatives with respect to n are given by

$$\begin{aligned}\log f(v|n) &= n \log \tilde{G}(v) + \log \Phi(v) + \log\left(1 - \frac{\tilde{G}(v-1)^n \Phi(v-1)}{\tilde{G}(v)^n \Phi(v)}\right) \\ &= n \log \tilde{G}(v) + \log\left(1 - r(v)^n \phi(v)\right)\end{aligned}$$

$$\frac{\partial}{\partial n}\log f(v|n) = \log \tilde{G}(v) - \frac{r(v)^n \log r(v)\phi(v)}{(1-r(v)^n\phi(v))}$$
$$= \log \tilde{G}(v) - \frac{\log r(v)\phi(v)}{(r(v)^{-n}-\phi(v))} \quad (38)$$
$$\frac{\partial^2}{\partial n^2}\log f(v|n) = \frac{r(v)^{-n}\log^2 r(v)\phi(v)}{(r(v)^{-n}-\phi(v))^2}.$$

The first derivative in equation 38 is often referred to as the score and denoted by $s_n(v)$. Since $0 < r(v) < 1$ and $\phi(v) > 0$, it is easy to verify that the composite log-marginal likelihood is strictly concave. □

D FIGURES

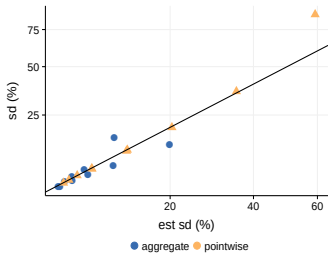

Figure 7: In the case with high overlap and $k_{small} = 10^6$, the std dev is correctly predicted despite bias in the estimates for the aggregate construction

E STREAMING AND COUNT-HLL-HIP ESTIMATORS

In streaming settings, the Historic Inverse Probability (HIP) estimator [7], [29] yields improved cardinality estimates by exploiting information available during sketch construction that are no longer available once the stream is completed. The HIP estimator augments a distinct count sketch with an additional counter. The counter is incremented by $1/p_t$ if the t^{th} distinct item actually changes the values in sketch. Here, p_t is the probability that the t^{th} item will update the value of the sketch given the sketch after time $t-1$.

We may augment the Count-HLL sketch in multiple ways. If there is a particular collection of labels of interest, then one may instantiate a counter for each of these labels. If the Count-HLL sketch is updated at time t by an item with a label ℓ of interest, then update the counter by $1/p_t(\ell)$ where $p_t(\ell)$ is the probability a random item added to label ℓ's registers will change the sketch at time t. If one wishes to estimate the variance as well, add additional counters which are incremented by $(1-p_t(\ell))/p_t(\ell)^2$ whenever the counts are incremented. These cardinality estimates are always unbiased, and the variance estimates are unbiased at the times where there is an increment [29]. Thus, the theoretical guarantees are also stronger than asymptotic guarantees.

If one is only interested in heavy hitters, that is the sets with exceptionally large cardinalities, then one can maintain a list of k counters corresponding to heavy hitters. For each item, label pair (x_t, ℓ_t), if the label corresponds to one of the heavy hitter counters, then update the counter as above in the fixed counter case. If the label does not correspond to an existing counter and the item changes the sketch, then estimate the cardinality from the sketch and compare it to the smallest heavy hitters counter. The new set replaces it if the estimated count is greater. This approach is similar how heavy hitters are maintained using the Count-Min sketch.

If one is interested in all counts, then one can store a separate CountMin sketch. The Count-HLL sketch produces a sequence of sets and increments $(\ell, 1/p_t(\ell))$. This $(label, count)$ sequence can be stored in a Count-Min sketch in the usual way.

Updates for Count-HLL-HIP sketches take $O(d)$ time compared to $O(1)$ for non-HIP sketches. When all sets have large cardinalities or when all entries in the sketch have many items hashed to them, this may not be problematic as there are logarithmically many updates.