

New cardinality estimation algorithms for HyperLogLog sketches

Draft version

Otmar Ertl
otmar.ertl@gmail.com

October 4, 2016

This paper presents new methods to estimate the cardinalities of multisets recorded by HyperLogLog sketches. A theoretically motivated extension to the original estimator is presented that eliminates the bias for small and large cardinalities. Based on the maximum likelihood principle another unbiased method is derived together with a robust and efficient numerical algorithm to calculate the estimate. The maximum likelihood method is also appropriate to improve cardinality estimates of set intersection sizes compared to the inclusion-exclusion principle. The new methods are demonstrated and verified by extensive simulations.

1. Introduction

Counting the number of distinct elements in a data stream or large datasets is a common problem in big data processing. Often there are parallel streams or data is spread over a cluster, which makes this task even more challenging. In principle, finding the number of distinct elements n with a maximum relative error ε in a data stream requires $\Omega(n)$ space [1]. However, probabilistic algorithms that achieve the requested accuracy only with high probability are able to drastically reduce space requirements. Many different probabilistic algorithms have been developed over the past two decades [2, 3]. An algorithm with an optimal space complexity of $\Omega(\varepsilon^{-2} + \log n)$ [1, 4] was finally presented [5]. This algorithm, however, is not very efficient in practice [3].

Currently, the most memory efficient algorithm that also works for distributed setups is the near-optimal HyperLogLog algorithm [6] with space complexity $\Omega(\varepsilon^{-2} \log \log n + \log n)$. The originally proposed estimation method has some problems to guarantee the

This paper together with source code for all presented algorithms and simulations is available at <https://github.com/oertl/hyperloglog-sketch-estimation-paper>.

same estimation error over the entire range of cardinalities. It was proposed to correct the estimate by empirical means [7, 8, 9].

In case the data is not distributed and results do not need to be aggregated further, there are even more efficient estimation algorithms available. On the one hand there is the self-learning bitmap [10, 11], and on the other hand there is the HyperLogLog algorithm extended by a historic inverse probability estimator that is continuously updated together with the HyperLogLog sketch [3]. Both achieve the same estimation error using less space. However, the estimated cardinality depends on the insertion order of elements and hence cannot be used in a distributed environment.

2. HyperLogLog data structure

The HyperLogLog algorithm uses a sketching data structure that consists of m registers. For performance reasons the number of registers m is chosen to be a power of 2, $m = 2^p$. p is the precision that directly influences the relative error which scales like $1/\sqrt{m}$. All registers start with zero initial value. Each element insertion potentially increases the value of one of these registers. The maximum value a register can reach is a natural bound given either by the output size of the used hash algorithm or the space that is reserved for a single register. Common implementations allocate up to 8 bits per register.

2.1. Data element insertion

In order to insert a data element into a HyperLogLog data structure a hash value is calculated. The leading p bits of the hash value are used to select one of the 2^p registers. Among the next q bits, the position of the first 1-bit is determined which is a value in the range $[1, q + 1]$. The value $q + 1$ is used, if all q bits are zeros. If the position of the first 1-bit exceeds the current value of the selected register, the register value is replaced. Algorithm 1 shows the update procedure for inserting a data element into the HyperLogLog sketch.

The described element insertion algorithm makes use of what is known as stochastic averaging [12]. Instead of updating each of all m registers using m independent hash values, which would be an $\mathcal{O}(m)$ operation, only one register is selected and updated, which requires only a single hash function and reduces the complexity to $\mathcal{O}(1)$.

A HyperLogLog sketch can be characterized by a parameter pair (p, q) . The first parameter p is the precision and controls the relative error while the second defines the possible value range of a register. A register can take all values starting from 0 to $q + 1$, inclusively. The sum $p + q$ corresponds to the number of consumed hash value bits and defines the maximum cardinality that can be tracked. Obviously, if the cardinality reaches values in the order of 2^{p+q} , hash collisions will become more apparent and the estimation accuracy will be drastically reduced.

At any time a (p, q) -HyperLogLog sketch with parameters (p, q) can be compressed into a (p', q') -HyperLogLog data structure, if $p' \leq p$ and $p' + q' \leq p + q$ is satisfied. This transformation is lossless in a sense that the resulting HyperLogLog sketch is the same

as if all elements would have been recorded by a (p', q') -HyperLogLog sketch right from the beginning.

Algorithm 1 has some properties which are especially useful in distributed environments. First, the insertion order of elements has no influence on the final HyperLogLog sketch. Furthermore, any two HyperLogLog sketches with same parameters (p, q) representing sets A and B can be easily merged. The resulting register values are simply given by the register-wise maximum values

$$K_i^{A \cup B} = \max(K_i^A, K_i^B) \quad \text{for } 1 \leq i \leq m. \quad (1)$$

The final HyperLogLog sketch represents the union of A and B .

A $(p, 0)$ -HyperLogLog sketch corresponds to a bit array as used by linear counting [13]. Each register value can be stored by a single bit in this case. Hence, linear counting can be regarded as a special case of the HyperLogLog algorithm for which $q = 0$.

Algorithm 1 Insertion of a data element D into a HyperLogLog data structure that consists of $m = 2^p$ registers. All registers $\mathbf{K} = (K_1, \dots, K_m)$ have zero initial value and remain in the range $[0, q + 1]$.

procedure INSERTELEMENT(D, \mathbf{K})
 $\langle a_1, \dots, a_p, b_1, \dots, b_q \rangle_2 \leftarrow (p + q)\text{-bit hash value of } D$
 $i \leftarrow 1 + \langle a_1, \dots, a_p \rangle_2 \quad \triangleright i \in [1, 2^p]$
 $k = \min(\{s \mid s \in [1, q] \wedge b_s = 1\} \cup \{q + 1\}) \quad \triangleright k \in [1, q + 1]$
 $K_i \leftarrow \max(K_i, k)$
end procedure

2.2. Joint probability distribution of register values

Under the assumption of a uniform hash function, the probability that the register values $\mathbf{K} = (K_1, \dots, K_m)$ of a HyperLogLog sketch with parameters p and q are equal to $\mathbf{k} = (k_1, \dots, k_m)$ is given by the corresponding probability mass function

$$\rho(\mathbf{k}|n) = \sum_{n_1 + \dots + n_m = n} \binom{n}{n_1, \dots, n_m} \frac{1}{m^n} \prod_{i=1}^m \gamma(k_i | n_i) \quad (2)$$

where n is the cardinality. The n distinct elements are distributed over all m registers according to a multinomial distribution with equal probabilities. $\gamma(k|n)$ is the probability that the value of a register is equal to k , after it was selected n times by the insertion algorithm

$$\gamma(k|n) := \begin{cases} 1 & n = 0 \wedge k = 0 \\ 0 & n = 0 \wedge 1 \leq k \leq q + 1 \\ 0 & n \geq 1 \wedge k = 0 \\ (1 - \frac{1}{2^k})^n - (1 - \frac{1}{2^{k-1}})^n & n \geq 1 \wedge 1 \leq k \leq q \\ 1 - (1 - \frac{1}{2^q})^n & n \geq 1 \wedge k = q + 1 \end{cases}. \quad (3)$$

The order of register values K_1, \dots, K_m is not important for the estimation of the cardinality. More formally, the multiset $\{K_1, \dots, K_m\}$ is a sufficient statistic for n . Since the values of the multiset are all in the range $[0, q+1]$ the multiset can also be represented as $\{K_1, \dots, K_m\} = 0^{C_0} 1^{C_1} \dots q^{C_q} (q+1)^{C_{q+1}}$ where C_k is the multiplicity of value k . As a consequence, the multiplicity vector $\mathbf{C} := (C_0, \dots, C_{q+1})$ is also a sufficient statistic for the cardinality. In addition, this vector contains all the information about the HyperLogLog sketch that is required for cardinality estimation. The two HyperLogLog parameters can be obtained by $p = \log_2 \|\mathbf{C}\|_1$ and $q = \dim \mathbf{C} - 2$, respectively.

2.3. Poisson approximation

Due to the statistical dependence of the register values, the probability mass function (2) is difficult for further analysis. Therefore, a Poisson model can be used [6], which assumes that the cardinality itself is distributed according to a Poisson distribution

$$n \sim \text{Poisson}(\lambda). \quad (4)$$

Under the Poisson model the distribution of the register values is

$$\begin{aligned} \rho(\mathbf{k}|\lambda) &= \sum_{n=0}^{\infty} \rho(\mathbf{k}|n) e^{-\lambda} \frac{\lambda^n}{n!} \\ &= \sum_{n_1=0}^{\infty} \dots \sum_{n_m=0}^{\infty} \prod_{i=1}^m \gamma(k_i|n_i) e^{-\frac{\lambda}{m}} \frac{\lambda^{n_i}}{n_i! m^{n_i}} \\ &= \prod_{i=1}^m \sum_{n=0}^{\infty} \gamma(k_i|n) e^{-\frac{\lambda}{m}} \frac{\lambda^n}{n! m^n} \\ &= \prod_{k=0}^{q+1} \left(\sum_{n=0}^{\infty} \gamma(k|n) e^{-\frac{\lambda}{m}} \frac{\lambda^n}{n! m^n} \right)^{c_k} \\ &= e^{-c_0 \frac{\lambda}{m}} \left(\prod_{k=1}^q \left(e^{-\frac{\lambda}{m 2^k}} \left(1 - e^{-\frac{\lambda}{m 2^k}} \right) \right)^{c_k} \right) \left(1 - e^{-\frac{\lambda}{m 2^q}} \right)^{c_{q+1}}. \end{aligned} \quad (5) \quad (6)$$

Here c_k denotes the multiplicity of value k in the multiset $\{k_1, \dots, k_m\}$. The final factorization shows that under the Poisson model the register values K_1, \dots, K_m are independent and identically distributed. The probability that a register has a value less than or equal to k for a given rate λ is defined by

$$P(K \leq k|\lambda) = \begin{cases} 0 & k < 0 \\ e^{-\frac{\lambda}{m 2^k}} & 0 \leq k \leq q \\ 1 & k > q \end{cases}. \quad (7)$$

2.4. Depoissonization

Due to the simpler probability mass function, it is easier to find an estimator $\hat{\lambda} = \hat{\lambda}(\mathbf{K})$ for the Poisson rate λ rather than for the cardinality n in the fixed-size model (2).

Depoissonization [14] finally allows to translate the estimates back to the fixed-size model. Assume we have found an unbiased estimator for the Poisson rate

$$\mathbb{E}(\hat{\lambda}|\lambda) = \lambda \quad \text{for all } \lambda \geq 0. \quad (8)$$

We know from (5)

$$\mathbb{E}(\hat{\lambda}|\lambda) = \sum_{n=0}^{\infty} \mathbb{E}(\hat{\lambda}|n) e^{-\lambda} \frac{\lambda^n}{n!} \quad (9)$$

and therefore

$$\sum_{n=0}^{\infty} \mathbb{E}(\hat{\lambda}|n) e^{-\lambda} \frac{\lambda^n}{n!} = \lambda \quad (10)$$

holds for all $\lambda \geq 0$. The unique solution of this equation is given by

$$\mathbb{E}(\hat{\lambda}|n) = n. \quad (11)$$

Hence, the unbiased estimator $\hat{\lambda}$ conditioned on n is also an unbiased estimator for n , which motivates us to use $\hat{\lambda}$ directly as estimator for the cardinality $\hat{n} := \hat{\lambda}$. As simulation results will show later, the Poisson approximation works well over the entire cardinality range, even for estimators that are not exactly unbiased.

3. Original cardinality estimation method

The original cardinality estimator [6] is based on the idea that the number of distinct element insertions a register needs to reach the value k is proportional to $m2^k$. Given that, a rough cardinality estimate can be obtained by calculating the average over the values $\{m2^{K_1}, \dots, m2^{K_m}\}$.

In the history of the HyperLogLog algorithm different averaging techniques have been proposed. First, there was the LogLog algorithm using the geometric mean and the SuperLogLog algorithm that enhanced the estimate by truncating the largest register values before applying the geometric mean [15]. Finally, the harmonic mean was found to give even better estimates as it is inherently less sensitive to outliers. The result is the so-called raw estimator which is given by

$$\hat{n}_{\text{raw}} = \alpha_m \frac{m}{\frac{1}{m2^{K_1}} + \dots + \frac{1}{m2^{K_m}}} = \frac{\alpha_m m^2}{\sum_{i=1}^m 2^{-K_i}} = \frac{\alpha_m m^2}{\sum_{k=0}^{q+1} C_k 2^{-k}}. \quad (12)$$

Here α_m is a bias correction factor which was derived for a given number of registers m to be [6]

$$\alpha_m := \left(m \int_0^{\infty} \left(\log_2 \left(\frac{2+u}{1+u} \right) \right)^m du \right)^{-1}. \quad (13)$$

Numerical approximations of α_m for various values of m have been listed in [6]. These approximations are used in many HyperLogLog implementations. However, since the published constants have been rounded to 4 significant digits, these approximations

even introduce some bias for very high precisions p . For HyperLogLog sketches that are used in practice with 256 or more registers ($p \geq 8$), it is completely sufficient to use

$$\alpha_\infty := \lim_{m \rightarrow \infty} \alpha_m = \frac{1}{2 \log 2} \approx 0.7213475, \quad (14)$$

as approximation for α_m in (12), because the additional bias is negligible compared to the overall estimation error.

Figure 1 shows the distribution of the relative error for the raw estimator as function of the cardinality. The chart is based on 10 000 randomly generated HyperLogLog sketches. More details of the experimental setup will be explained later in Section 3.5. Obviously, the raw estimator is biased for small and large cardinalities where it fails to return accurate estimates. In order to cover the entire range of cardinalities, corrections for small and large cardinalities have been proposed.

As mentioned in Section 2.1, a HyperLogLog sketch with parameters (p, q) can be mapped to a $(p, 0)$ -HyperLogLog sketch. Since $q = 0$ corresponds to linear counting and the reduced HyperLogLog sketch corresponds to a bitset with C_0 zeros, the linear counting cardinality estimator [13] can be used

$$\hat{n}_{\text{small}} = m \log(m/C_0). \quad (15)$$

The corresponding relative estimation error as depicted in Figure 2 shows that this estimator is convenient for small cardinalities. It was proposed to use this estimator for small cardinalities as long as $\hat{n}_{\text{raw}} \leq \frac{5}{2}m$ where the factor $\frac{5}{2}$ was empirically determined [6].

For large cardinalities in the order of 2^{p+q} , for which a lot of registers are already in a saturated state, meaning that they have reached the maximum possible value $q + 1$, the raw estimator underestimates the cardinalities. For the 32-bit hash value case ($p + q = 32$), which was considered in [6], following correction formula was proposed to take these saturated registers into account

$$\hat{n}_{\text{large}} = -2^{32} \log(1 - \hat{n}_{\text{raw}}/2^{32}). \quad (16)$$

The original estimation algorithm as presented in [6] including corrections for small and large cardinalities is summarized by Algorithm 2. The relative estimation error for the original method is shown in Figure 3. Unfortunately, as can be clearly seen, the ranges where the estimation error is small for \hat{n}_{raw} and \hat{n}_{small} do not overlap. Therefore, the estimation error is much larger near the transition region. To reduce the estimation error for cardinalities close to this region, it was proposed to correct \hat{n}_{raw} for bias. Empirically collected bias correction data is either stored as set of interpolation points [7], as lookup table [8], or as best-fitting polynomial [9]. However, all these empirical approaches treat the symptom and not the cause.

The large range correction formula (16) is not satisfying either as it does not reduce the estimation error. Quite the contrary, it even makes the bias worse. However, instead of underestimating the cardinalities, they are now overestimated. Another indication for the incorrectness of the proposed large range correction is the fact that it is not even

Algorithm 2 Original cardinality estimation algorithm for HyperLogLog sketches that use 32-bit hash values ($p + q = 32$) for insertion of data items [6].

```

function ESTIMATECARDINALITY( $\mathbf{K}$ )
   $m \leftarrow \dim \mathbf{K}$ 
   $\hat{n}_{\text{raw}} = \alpha_m m^2 (\sum_{i=1}^m 2^{-K_i})^{-1}$  ▷ raw estimate (12)
  if  $\hat{n}_{\text{raw}} \leq \frac{5}{2}m$  then
     $C_0 = |\{i | K_i = 0\}|$ 
    if  $C_0 \neq 0$  then
      return  $m \log(m/C_0)$  ▷ small range correction (15)
    else
      return  $\hat{n}_{\text{raw}}$ 
    end if
  else if  $\hat{n}_{\text{raw}} \leq \frac{1}{30}2^{32}$  then
    return  $\hat{n}_{\text{raw}}$ 
  else
    return  $-2^{32} \log(1 - \hat{n}_{\text{raw}}/2^{32})$  ▷ large range correction (16)
  end if
end function

```

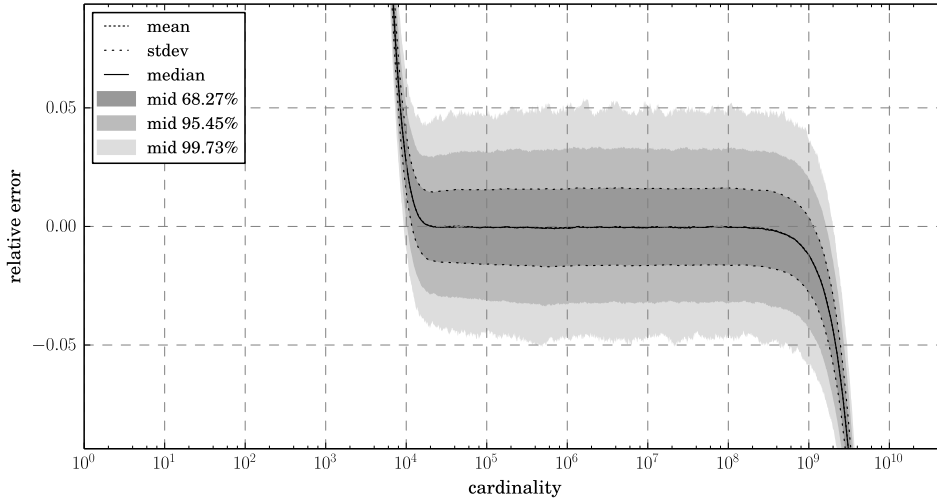


Figure 1: The distribution of the relative estimation error over the cardinality for the raw estimator after evaluation of 10 000 randomly generated HyperLogLog data structures with parameters $p = 12$ and $q = 20$.

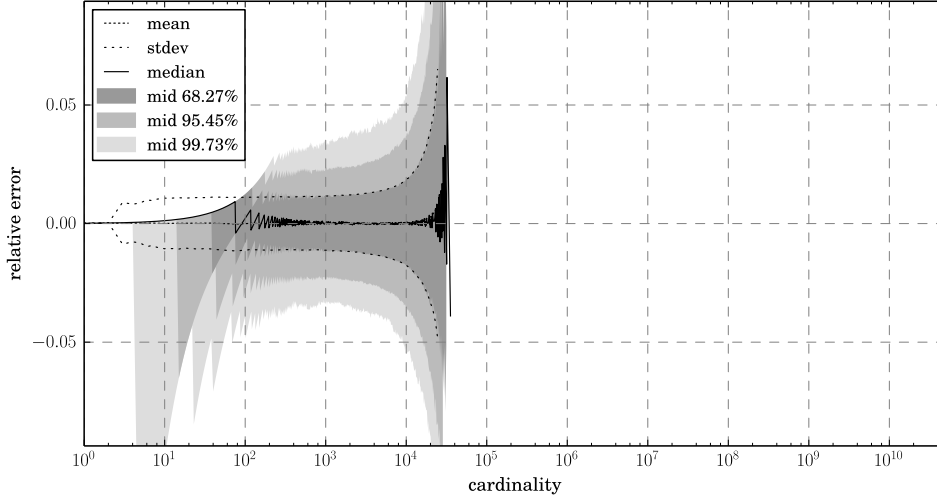


Figure 2: The distribution of the relative estimation error for the linear counting estimator after evaluation of 10 000 randomly generated bitmaps of size 2^{12} which correspond to HyperLogLog sketches with parameters $p = 12$ and $q = 0$.

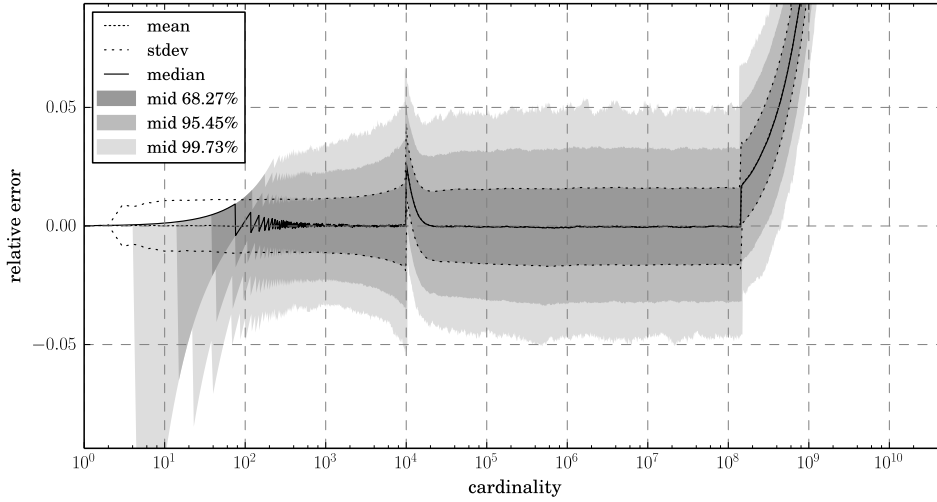


Figure 3: The distribution of the relative estimation error over the cardinality for the original estimation algorithm after evaluation of 10 000 randomly generated HyperLogLog data structures with parameters $p = 12$ and $q = 20$.

defined for all possible states. For instance, consider a (p, q) -HyperLogLog sketch with $p + q = 32$ for which all registers are equal to the maximum possible value $q + 1$. The raw estimate would be $\hat{n}_{\text{raw}} = \alpha_m 2^{33}$, which is greater than 2^{32} and outside of the domain of the large range correction formula.

A simple approach to avoid the need of any large range correction is to extend the operating range of the raw estimator to larger cardinalities. This can be easily accomplished by using hash values with more bits ($p + q > 32$). If $q \geq 30$, 5-bit registers are no longer sufficient to represent all possible values in the range $[0, q + 1]$. If, for example, 64-bit hash values ($p + q = 64$) are used, as proposed in [7], 6 bits per register are needed. Hash value sizes with more than 64 bits are useless in practice, because it is unrealistic to encounter cardinalities of order 2^{64} for which the raw estimator would be biased.

3.1. Derivation of the raw estimator

In order to better understand why the raw estimator fails for small and large cardinalities, we start with a brief and simple derivation without the restriction to large cardinalities ($n \rightarrow \infty$) and without using complex analysis as in [6].

Let us assume that the register values have following cumulative distribution function

$$P(K \leq k | \lambda) = e^{-\frac{\lambda}{m2^k}}. \quad (17)$$

For now we ignore that this distribution has infinite support and differs from the register value distribution under the Poisson model (7), whose support is limited to the range $[0, q + 1]$. For a random variable K obeying (17) the expectation of 2^{-K} is given by

$$\begin{aligned} \mathbb{E}(2^{-K}) &= \\ \sum_{k=-\infty}^{\infty} 2^{-k} \left(e^{-\frac{\lambda}{m2^k}} - e^{-\frac{\lambda}{m2^{k+1}}} \right) &= \frac{1}{2} \sum_{k=-\infty}^{\infty} 2^k e^{-\frac{\lambda}{m} 2^k} = \frac{\alpha_{\infty} m \xi(\log_2(\lambda/m))}{\lambda}, \end{aligned} \quad (18)$$

where the function

$$\xi(x) := \log(2) \sum_{k=-\infty}^{\infty} 2^{k+x} e^{-2^{k+x}} \quad (19)$$

is a smooth periodic function with period 1. Numerical evaluations indicate that this function can be bounded by $1 - \varepsilon_{\xi} \leq \xi(x) \leq 1 + \varepsilon_{\xi}$ with $\varepsilon_{\xi} := 9.885 \times 10^{-6}$ (see Figure 4).

Let K_1, \dots, K_m be a sample taken from (17). For large sample sizes $m \rightarrow \infty$ we have asymptotically

$$\mathbb{E} \left(\frac{1}{2^{-K_1} + \dots + 2^{-K_m}} \right) \stackrel{m \rightarrow \infty}{=} \frac{1}{\mathbb{E}(2^{-K_1} + \dots + 2^{-K_m})} = \frac{1}{m \mathbb{E}(2^{-K})}. \quad (20)$$

Together with (18) we obtain

$$\lambda = \mathbb{E} \left(\frac{\alpha_{\infty} m^2 \xi(\log_2(\lambda/m))}{2^{-K_1} + \dots + 2^{-K_m}} \right) \quad \text{for } m \rightarrow \infty. \quad (21)$$



Figure 4: The deviation of $\xi(x)$ from 1.

Therefore, the asymptotic relative bias of

$$\hat{\lambda} = \frac{\alpha_{\infty} m^2}{2^{-K_1} + \dots + 2^{-K_m}} \quad (22)$$

is bounded by ε_{ξ} , which makes this statistic a good estimator for the Poisson parameter. It also corresponds to the raw estimator (12), if the Poisson parameter estimate is used as cardinality estimate (see Section 2.4).

3.2. Limitations of the raw estimator

The raw estimator is based on two prerequisites. First, the number of registers needs to be sufficiently large ($m \rightarrow \infty$). And second, the distribution of register values can be described by (17). However, the latter is not true for small and large cardinalities, which is finally the reason for the bias of the raw estimator.

A random variable K' with cumulative distribution function (17) can be transformed into a random variable K with cumulative distribution function (7) using

$$K = \min(\max(K', 0), q + 1). \quad (23)$$

Therefore, the register values K_1, \dots, K_m can be seen as the result after applying this transformation to a sample K'_1, \dots, K'_m of the distribution described by (17). If all registers values of the HyperLogLog sketch fall into the range $[1, q]$, they must be identical to the values K'_1, \dots, K'_m . In other words, the observed register values are also a plausible sample of the assumed distribution described by (17) in this case. Hence, as long as all or at least most register values are in the range $[1, q]$, the approximation of (7) by (17) is valid. This explains why the raw estimator works best for intermediate cardinalities. However, for small and large cardinalities many registers have values equal to 0 or $q + 1$, respectively, which cannot be described by (17) and finally leads to the observed bias.

3.3. Corrections to the raw estimator

If we knew the values K'_1, \dots, K'_m for which the transformation (23) led to the observed register values K_1, \dots, K_m , we would be able to estimate λ using

$$\hat{\lambda} = \frac{\alpha_\infty m^2}{2^{-K'_1} + \dots + 2^{-K'_m}}. \quad (24)$$

As we have already shown, this estimator is approximately unbiased, because all K'_i follow the assumed distribution. It would be even sufficient, if we knew the multiplicity C'_k of each $k \in \mathbb{Z}$ in $\{K'_1, \dots, K'_m\}$, $C'_k := |\{i | k = K'_i\}|$, because the raw estimator can be also written as

$$\hat{\lambda} = \frac{\alpha_\infty m^2}{\sum_{k=-\infty}^{\infty} C'_k 2^{-k}}. \quad (25)$$

Due to (23), the multiplicities C'_k and the multiplicities C_k for the observed register values have following relationships

$$\begin{aligned} C_0 &= \sum_{k=-\infty}^0 C'_k, \\ C_k &= C'_k, \quad 1 \leq k \leq q, \\ C_{q+1} &= \sum_{k=q+1}^{\infty} C'_k. \end{aligned} \quad (26)$$

The idea is now to find estimates \hat{c}'_k for all $k \in \mathbb{Z}$ and use them as replacements for C'_k in (25). For $k \in [1, q]$ where $C_k = C'_k$ we can use the trivial estimators

$$\hat{c}'_k := C_k, \quad 1 \leq k \leq q. \quad (27)$$

To get estimators for $k \leq 0$ and $k \geq q+1$, we consider the expectation of C'_k

$$\mathbb{E}(C'_k) = m e^{-\frac{\lambda}{m^{2k}}} \left(1 - e^{-\frac{\lambda}{m^{2q}}}\right). \quad (28)$$

From (7) we know that $\mathbb{E}(C_0/m) = e^{-\frac{\lambda}{m}}$ and $\mathbb{E}(1 - C_{q+1}/m) = e^{-\frac{\lambda}{m^{2q}}}$, and therefore, we can also write

$$\mathbb{E}(C'_k) = m (\mathbb{E}(C_0/m))^{2^{-k}} \left(1 - (\mathbb{E}(C_0/m))^{2^{-k}}\right) \quad (29)$$

and

$$\mathbb{E}(C'_k) = m (\mathbb{E}(1 - C_{q+1}/m))^{2^{q-k}} \left(1 - (\mathbb{E}(1 - C_{q+1}/m))^{2^{q-k}}\right), \quad (30)$$

which motivates us to use

$$\hat{c}'_k = m (C_0/m)^{2^{-k}} \left(1 - (C_0/m)^{2^{-k}}\right) \quad (31)$$

as estimator for $k \leq 0$ and

$$\hat{c}'_k = m (1 - C_{q+1}/m)^{2^{q-k}} \left(1 - (1 - C_{q+1}/m)^{2^{q-k}}\right) \quad (32)$$

as estimator for $k \geq q + 1$, respectively.

Inserting all these estimators into (25) as replacements for C'_k finally gives

$$\hat{\lambda} = \frac{\alpha_\infty m^2}{\sum_{k=-\infty}^{\infty} \hat{c}'_k 2^{-k}} = \frac{\alpha_\infty m^2}{m \sigma(C_0/m) + \sum_{k=1}^q C_k 2^{-k} + m \tau(1 - C_{q+1}/m) 2^{-q}} \quad (33)$$

which we call the corrected raw estimator. Here $m \sigma(C_0/m)$ and $2m \tau(1 - C_{q+1}/m)$ are replacements for C_0 and C_{q+1} in the raw estimator (12), respectively. The functions σ and τ are defined as

$$\sigma(x) := x + \sum_{k=1}^{\infty} x^{2^k} 2^{k-1} \quad (34)$$

and

$$\tau(x) := \sum_{k=1}^{\infty} x^{2^{-k}} \left(1 - x^{2^{-k}}\right) 2^{-k}. \quad (35)$$

Using the identity $\sigma(x) + \tau(x) = \alpha_\infty \xi(\log_2(\log(1/x))) / \log(1/x)$, we get for the linear counting case with $q = 0$

$$\hat{\lambda} = \frac{\alpha_\infty m}{\sigma(C_0/m) + \tau(C_0/m)} = \frac{m \log(m/C_0)}{\xi(\log_2(\log(m/C_0)))} \approx m \log(m/C_0), \quad (36)$$

which is almost identical to the linear counting estimator used for the small range correction (15). Here we used again the fact that the function ξ can be well approximated by 1 (see Section 3.1).

3.4. Corrected raw estimation algorithm

The corrected raw estimator (33) leads to a new cardinality estimation algorithm for HyperLogLog sketches. Algorithm 3 demonstrates the numerical evaluation of the estimator. The values of functions σ and τ can be either calculated on-demand or precalculated. The possible value range of C_0 and C_{q+1} is $[0, m]$. Therefore, if performance matters, the function values can be precalculated for all possible arguments and kept in lookup tables of size $m + 1$.

3.5. Estimation error

In order to verify the new estimation algorithm, we generated 10 000 HyperLogLog sketches and inserted up to 50 billion unique elements. Assuming a uniform hash function, element hash values can be mocked by random numbers. For the following results we used the Mersenne Twister random number generator with 19 937 bit state size from the C++ standard library.

Figure 5 shows the distribution of the relative error of the estimated cardinality using Algorithm 3 compared to the true cardinality for $p = 12$ and $q = 20$. As the mean shows, the error is unbiased over the entire cardinality range. The new approach is able to accurately estimate cardinalities up to 4 billions ($\approx 2^{p+q}$) which is about an order of

Algorithm 3 Cardinality estimation algorithm based on the corrected raw estimator.

```

function ESTIMATECARDINALITY( $\mathbf{C}$ )
   $m \leftarrow \|\mathbf{C}\|_1$ 
   $z \leftarrow m \cdot \tau(1 - C_{q+1}/m)$   $\triangleright$  alternatively, take  $m \cdot \tau(1 - C_{q+1}/m)$  from
    precalculated lookup table

  for  $k \leftarrow q, 1$  do
     $z \leftarrow 0.5 \cdot (z + C_k)$ 
  end for
   $z \leftarrow z + m \cdot \sigma(C_0/m)$   $\triangleright$  alternatively, take  $m \cdot \sigma(C_0/m)$  from pre-
    calculated lookup table

  return  $\alpha_\infty m^2 / z$   $\triangleright \alpha_\infty := 1/(2 \log(2))$ 
end function

function  $\sigma(x)$   $\triangleright x \in [0, 1]$ 
  if  $x = 1$  then
    return  $\infty$ 
  end if
   $y \leftarrow 1$ 
   $z \leftarrow x$ 
  repeat
     $x \leftarrow x \cdot x$ 
     $z' \leftarrow z$ 
     $z \leftarrow z + x \cdot y$ 
     $y \leftarrow 2 \cdot y$ 
  until  $z = z'$ 
  return  $z$ 
end function

function  $\tau(x)$   $\triangleright x \in [0, 1]$ 
   $y \leftarrow 1$ 
   $z \leftarrow 0$ 
  repeat
     $x \leftarrow \sqrt{x}$ 
     $z' \leftarrow z$ 
     $y \leftarrow 0.5 \cdot y$ 
     $z \leftarrow z + (1 - x) \cdot x \cdot y$ 
  until  $z = z'$ 
  return  $z$ 
end function

```

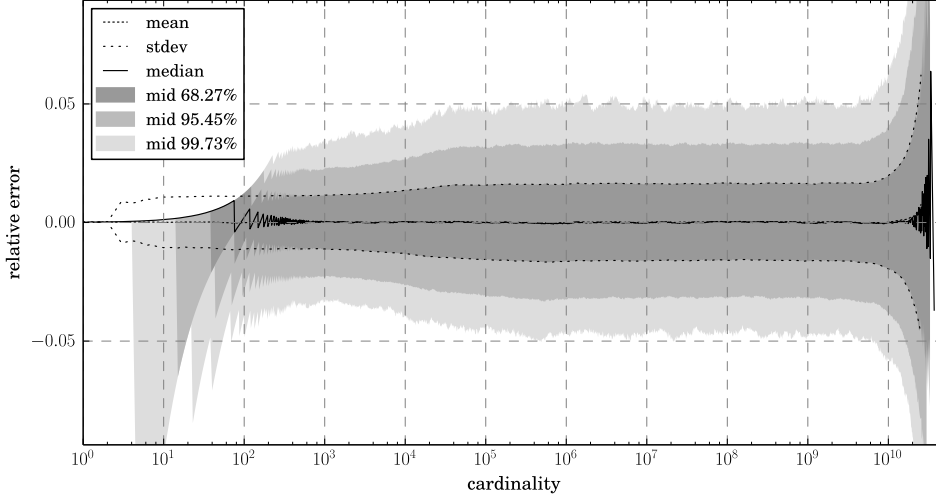


Figure 5: Relative error of the corrected raw estimator as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 12$ and $q = 20$.

magnitude larger than the operating range upper bound of the raw estimator (Figure 1) or the original method (Figure 3).

The corrected raw estimator beats the precision of methods that use bias correction [7, 8, 9]. Figure 6 shows the standard deviation of the raw estimator after bias correction. At the transition region to small cardinalities, for which all previous methods switch over to the linear counting estimator at some point, the corrected raw estimator has smaller variance. Obviously, the previous approaches cannot do better than given by the minimum of both curves for linear counting and raw estimator. In practice, the standard deviation is even larger, because the choice between both estimators must be made based on an estimate and not on the true cardinality, for which the intersection point of both curves represents the ideal transition point.

The new estimation algorithm also works well for other HyperLogLog configurations. First we considered configurations using a 32-bit hash function ($p + q = 32$). The relative estimation error for precisions $p = 8$, $p = 16$, $p = 22$ are shown in Figure 7, Figure 8, and Figure 9, respectively. As expected, since $p + q = 32$ is kept constant, the operating range remains more or less the same, while the relative error decreases with increasing precision. Again, the new algorithm gives essentially unbiased estimates. Only for very large precisions, an oscillating bias becomes apparent (compare Figure 9), that is caused by approximating the periodic function ξ by a constant (see Section 3.1).

As proposed in [7], the operating range can be extended by replacing the 32-bit hash function by a 64-bit hash function. Figure 10 shows the relative error for such a HyperLogLog configuration with parameters $p = 12$ and $q = 52$. The doubled hash value size shifts the maximum trackable cardinality value towards 2^{64} . As Figure 10 shows, when

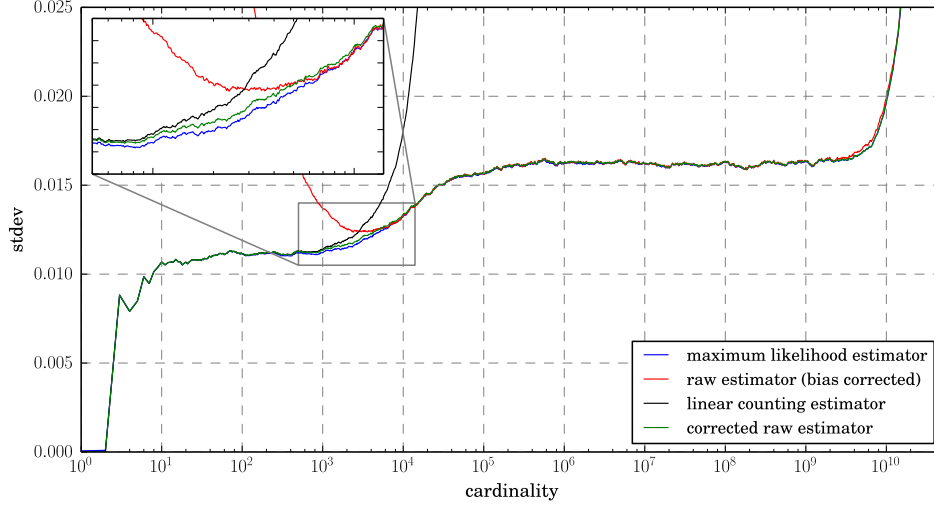


Figure 6: Standard deviations of different cardinality estimators for a HyperLogLog sketch with parameters $p = 12$ and $q = 20$ as functions of the true cardinality.

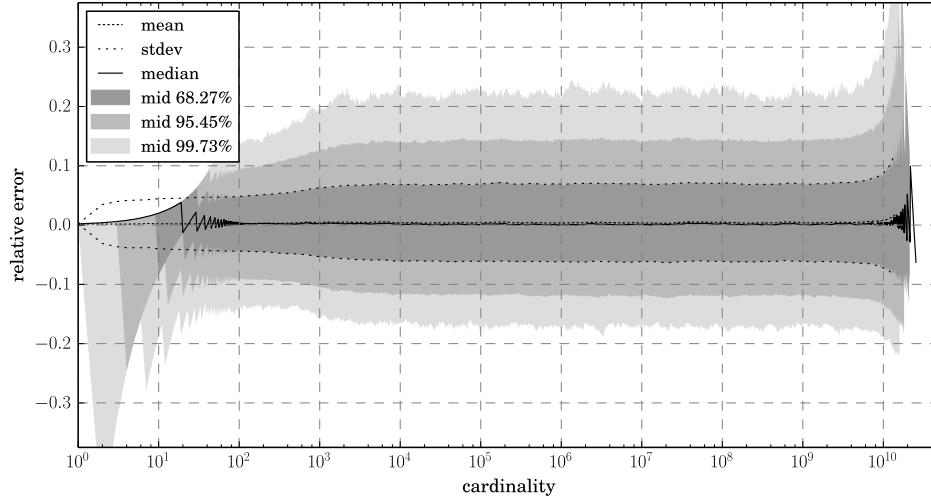


Figure 7: Relative error of the corrected raw estimator as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 8$ and $q = 24$.

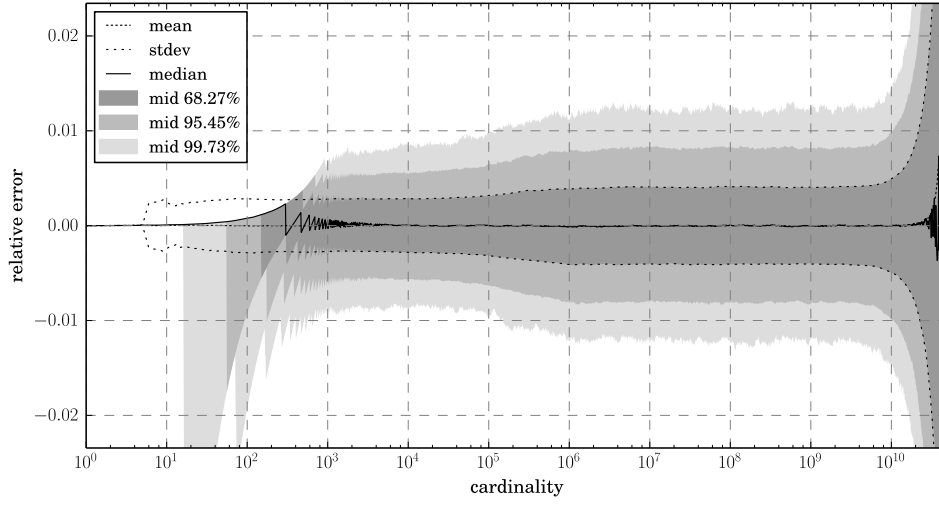


Figure 8: Relative error of the corrected raw estimator as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 16$ and $q = 16$.

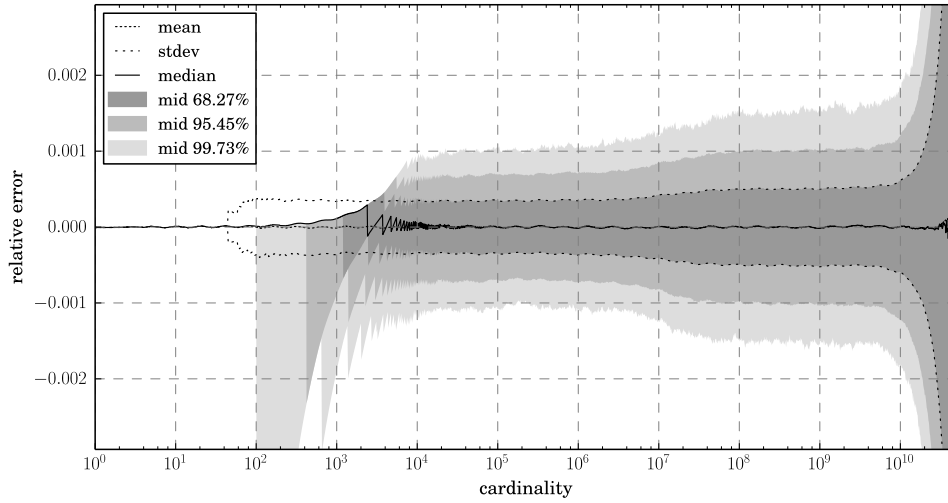


Figure 9: Relative error of the corrected raw estimator as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 22$ and $q = 10$.

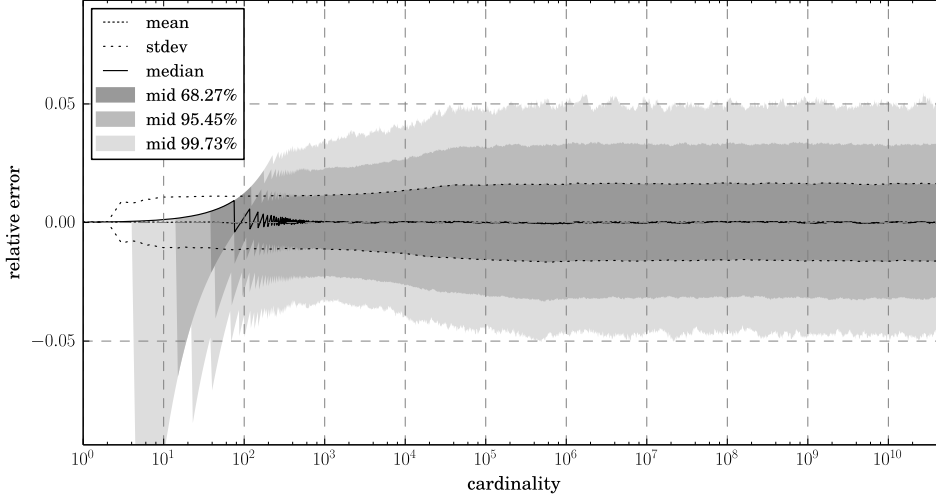


Figure 10: Relative error of the corrected raw estimator as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 12$ and $q = 52$.

compared to the 32-bit hash value case given in Figure 5, the estimation error remains constant over the entire simulated cardinality range up to 50 billions.

We also evaluated the case $p = 12$ and $q = 14$, which is interesting, because the register values are limited to the range $[0, 15]$. As a consequence, 4 bits are sufficient for representing a single register value. This allows two registers to share a single byte, which is beneficial from a performance perspective. Nevertheless, this configuration still allows the estimation of cardinalities up to 100 millions as shown in Figure 11, which could be enough for many applications.

3.6. Performance

In order to evaluate the performance of the corrected raw estimation algorithm, we investigated the average computation time for estimating the cardinality from a given HyperLogLog sketch. For different cardinalities we loaded precalculated multiplicity vectors of 1000 randomly generated HyperLogLog sketches into main memory. The average computation time was determined by cycling over these multiplicity vectors and passing them as input to the algorithm. For each evaluated cardinality value the average execution time was calculated after 100 cycles which corresponds to 100 000 algorithm executions for each cardinality value. The results for HyperLogLog configurations $p = 12, q = 20$ and $p = 12, q = 52$ are shown in Figure 12. Two variants of Algorithm 3 have been evaluated for which the functions σ and τ have been either calculated on-demand or taken from a lookup table. All these benchmarks were carried out on an Intel Core i5-2500K clocking at 3.3 GHz.

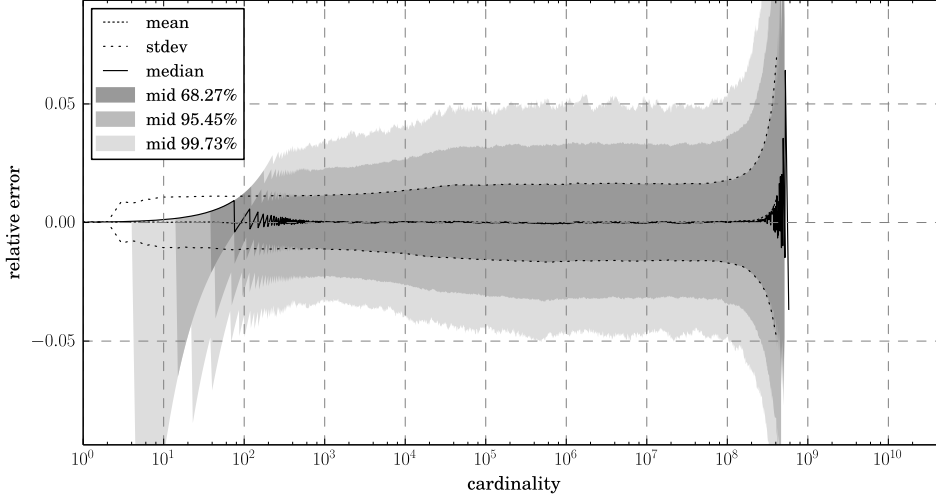


Figure 11: Relative error of the corrected raw estimator as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 12$ and $q = 14$.

The results show that the execution times are nearly constant for $q = 52$. Using a lookup table makes not much difference, because the on-demand calculation for σ is very fast and the calculation of τ is rarely needed due to the small probability of saturated registers, $P(C_{53} > 0) \approx 0$, for realistic cardinalities.

For the case $q = 20$ with precalculated correction values the computation time is again – as expected – independent of the cardinality. The faster computation times compared to the $q = 52$ case can be explained by the much smaller dimension of the multiplicity vector which is equal to $q + 2$. If the functions σ and τ are calculated on-demand, the execution times for larger cardinalities are doubled. This comes from the fact that the calculation of τ is much more expensive than that of σ , because it requires more iterations and involves square root evaluations. However, we can imagine that a better numerical approximation of τ can be found than that given in Algorithm 3 and which allows on-demand evaluation without significant extra costs.

The numbers do not yet include the required processing time to extract the multiplicity vector out from the HyperLogLog sketch, which requires a complete scan over all registers and counting the different register values into an array. A theoretical lower bound for this processing time can be derived using the maximum memory bandwidth of the CPU, which is 21 GB/s for an Intel Core i5-2500K. If we consider a HyperLogLog sketch with precision $p = 12$ which uses 5 bits per register, the total data size of the HyperLogLog sketch is 2.5 kB minimum. Consequently, the transfer time from main memory to CPU will be at least 120 ns. Having this value in mind, the presented numbers for estimating the cardinality from the multiplicity vector are quite satisfying.

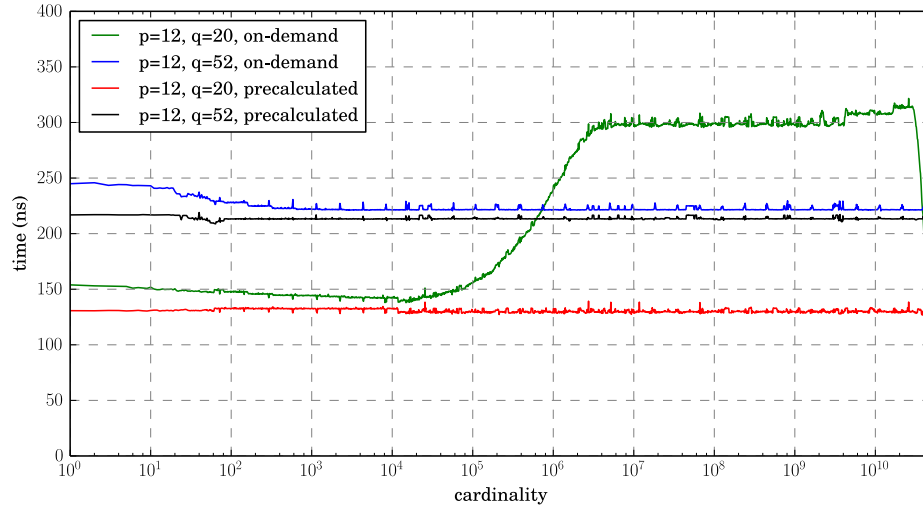


Figure 12: Average computation time as a function of the true cardinality with an Intel Core i5-2500K clocking at 3.3 GHz when estimating the cardinality from HyperLogLog sketches with parameters $p = 12$, $q = 20$ and $p = 12$, $q = 52$, respectively. Both cases, σ and τ precalculated and calculated on-demand have been considered.

4. Maximum likelihood estimation

We know from Section 2.4 that any unbiased estimator for the Poisson parameter is also an unbiased estimator for the cardinality. Moreover, we know that under suitable regularity conditions of the probability mass function the maximum likelihood estimator is asymptotically efficient [16]. This means, if the number of registers m is large enough, the maximum likelihood method should give us an unbiased estimator for the cardinality.

We remark that we are not the first applying the maximum likelihood method to HyperLogLog sketches [17], where the case without stochastic averaging (compare Section 2.1) was considered. The register values are statistically independent by nature in this case which allows factorization of the joint probability mass function and allows a straightforward application of the maximum likelihood method.

The more relevant case with stochastic averaging, which is considered in this paper, is more complicated. However, with the help of the Poisson approximation, we are able to still apply the maximum likelihood method and to derive a new robust and efficient cardinality estimation algorithm. Furthermore, we will demonstrate that consequent application of the maximum likelihood method reveals that the cardinality estimate needs to be roughly proportional to the harmonic mean for intermediate cardinality values. The history of the HyperLogLog algorithm shows that the raw estimator (12) was first found after several attempts using the geometric mean [6, 15].

4.1. Log-likelihood function

Using the probability mass function of the Poisson model (6) the log-likelihood and its derivative are given by

$$\log \mathcal{L}(\lambda|\mathbf{K}) = -\frac{\lambda}{m} \sum_{k=0}^q \frac{C_k}{2^k} + \sum_{k=1}^q C_k \log\left(1 - e^{-\frac{\lambda}{m2^k}}\right) + C_{q+1} \log\left(1 - e^{-\frac{\lambda}{m2^q}}\right) \quad (37)$$

and

$$\frac{d}{d\lambda} \log \mathcal{L}(\lambda|\mathbf{K}) = -\frac{1}{\lambda} \left(\frac{\lambda}{m} \sum_{k=0}^q \frac{C_k}{2^k} + \sum_{k=1}^q C_k \frac{\frac{\lambda}{m2^k}}{1 - e^{-\frac{\lambda}{m2^k}}} + C_{q+1} \frac{\frac{\lambda}{m2^q}}{1 - e^{-\frac{\lambda}{m2^q}}} \right). \quad (38)$$

As a consequence, the maximum likelihood estimate for the Poisson parameter is given by

$$\hat{\lambda} = m\hat{x}, \quad (39)$$

if \hat{x} is the root of the function

$$f(x) := x \sum_{k=0}^q \frac{C_k}{2^k} + \sum_{k=1}^q C_k \frac{\frac{x}{2^k}}{1 - e^{-\frac{x}{2^k}}} + C_{q+1} \frac{\frac{x}{2^q}}{1 - e^{-\frac{x}{2^q}}}. \quad (40)$$

This function can also be written as

$$f(x) := x \sum_{k=0}^q \frac{C_k}{2^k} + \sum_{k=1}^q C_k h\left(\frac{x}{2^k}\right) + C_{q+1} h\left(\frac{x}{2^q}\right) - (m - C_0), \quad (41)$$



Figure 13: The function $h(x)$.

where the function $h(x)$ is defined as

$$h(x) := 1 - \frac{x}{e^x - 1}. \quad (42)$$

$h(x)$ is strictly increasing and concave as can be seen in Figure 13. For non-negative values x the function ranges from $h(0) = 0$ to $h(x \rightarrow \infty) = 1$. Since the function $f(x)$ is also strictly increasing, it is obvious that there exists a unique root \hat{x} for which $f(\hat{x}) = 0$. The function is non-positive at 0 since $f(0) = C_0 - m \leq 0$ and, in case $C_{q+1} < m$ which implies $\sum_{k=0}^q \frac{C_k}{2^k} > 0$, the function is at least linearly increasing. $C_{q+1} = m$ corresponds to the case with all registers equal to the maximum value $q + 1$, for which the maximum likelihood estimate would be infinite.

It is easy to see that the estimate $\hat{\lambda}$ remains equal or becomes larger, when inserting an element into the HyperLogLog sketch following Algorithm 1. An update potentially changes the multiplicity vector (C_0, \dots, C_{q+1}) to $(C_0, \dots, C_i - 1, \dots, C_j + 1, \dots, C_{q+1})$ where $i < j$. Writing (41) as

$$f(x) := C_0 x + C_1 \left(h\left(\frac{x}{2^1}\right) + \frac{x}{2^1} - 1 \right) + C_2 \left(h\left(\frac{x}{2^2}\right) + \frac{x}{2^2} - 1 \right) + \dots \\ \dots + C_q \left(h\left(\frac{x}{2^q}\right) + \frac{x}{2^q} - 1 \right) + C_{q+1} \left(h\left(\frac{x}{2^q}\right) - 1 \right). \quad (43)$$

shows that the coefficient of C_i is larger than the coefficient of C_j in case $i < j$. Keeping x fixed during an update decreases $f(x)$. As a consequence, since $f(x)$ is increasing, the new root and hence the estimate must be larger than prior the update.

For the special case $q = 0$, which corresponds to the already mentioned linear counting algorithm, (40) can be solved analytically. In this case, the maximum likelihood method under the Poisson model leads directly to the linear counting estimator (15). Due to this fact we could expect that maximum likelihood estimation under the Poisson model also works well for the more general HyperLogLog case.

4.2. Inequalities for the maximum likelihood estimate

In the following lower and upper bounds for \hat{x} are derived. Applying Jensen's inequality on h in (41) gives an upper bound for $f(x)$:

$$f(x) \leq x \sum_{k=0}^q \frac{C_k}{2^k} + (m - C_0) \cdot h\left(x \cdot \frac{\sum_{k=1}^q \frac{C_k}{2^k} + \frac{C_{q+1}}{2^q}}{m - C_0}\right) - (m - C_0). \quad (44)$$

The left-hand side is zero, if \hat{x} is inserted. Resolution for \hat{x} finally gives the lower bound

$$\hat{x} \geq \frac{m - C_0}{\sum_{k=1}^q \frac{C_k}{2^k} + \frac{C_{q+1}}{2^q}} \log\left(1 + \frac{\sum_{k=1}^q \frac{C_k}{2^k} + \frac{C_{q+1}}{2^q}}{\sum_{k=0}^q \frac{C_k}{2^k}}\right). \quad (45)$$

This bound can be weakened using $\log(1+x) \geq \frac{2x}{x+2}$ for $x \geq 0$ which results in

$$\hat{x} \geq \frac{m - C_0}{C_0 + \frac{3}{2} \sum_{k=1}^q \frac{C_k}{2^k} + \frac{C_{q+1}}{2^q}}. \quad (46)$$

Using the monotonicity of h , the lower bound

$$f(x) \geq x \sum_{k=0}^q \frac{C_k}{2^k} + \sum_{k=1}^q C_k h\left(\frac{x}{2^{K'_{\max}}}\right) + C_{q+1} h\left(\frac{x}{2^{K'_{\max}}}\right) - (m - C_0) \quad (47)$$

can be found, where $K'_{\max} := \min(K_{\max}, q)$ and $K_{\max} := \max\{k | C_k > 0\}$. Again, inserting \hat{x} and transformation gives

$$\hat{x} \leq 2^{K'_{\max}} \log\left(1 + \frac{m - C_0}{2^{K'_{\max}} \sum_{k=0}^q \frac{C_k}{2^k}}\right) \quad (48)$$

as upper bound which can be weakened using $\log(1+x) \leq x$ for $x \geq 0$

$$\hat{x} \leq \frac{m - C_0}{\sum_{k=0}^q \frac{C_k}{2^k}}. \quad (49)$$

If the HyperLogLog sketch is in the intermediate range, where $C_0 = C_{q+1} = 0$ the bounds (46) and (49) differ only by a constant factor from the raw estimator (12). Hence, the maximum likelihood method leads directly to the harmonic mean that is used by the raw estimator.

4.3. Computation of the maximum likelihood estimate

Since f is concave and increasing, both, Newton-Raphson iteration and the secant method, will converge to the root, provided that the function is negative for the starting points. In the following we start from the secant method to derive the new cardinality estimation algorithm. Even though the secant method has the disadvantage of slower

convergence, a single iteration is simpler to calculate as it does not require the evaluation of the first derivative. An iteration step of the secant method can be written as

$$x_i = x_{i-1} - (x_{i-1} - x_{i-2}) \frac{f(x_{i-1})}{f(x_{i-1}) - f(x_{i-2})} \quad (50)$$

If we set x_0 equal to 0, for which $f(x_0) = -(m - C_0)$, and x_1 equal to one of the derived lower bounds (45) or (46), the sequence (x_0, x_1, x_2, \dots) is montone increasing. Using the definitions

$$\Delta x_i := x_i - x_{i-1} \quad (51)$$

and

$$g(x) := f(x) + (m - C_0) = x \sum_{k=0}^q \frac{C_k}{2^k} + \sum_{k=1}^q C_k h\left(\frac{x}{2^k}\right) + C_{q+1} h\left(\frac{x}{2^q}\right) \quad (52)$$

the iteration scheme can also be written as

$$\Delta x_i = \Delta x_{i-1} \frac{(m - C_0) - g(x_{i-1})}{g(x_{i-1}) - g(x_{i-2})} \quad (53)$$

$$x_i = x_{i-1} + \Delta x_i \quad (54)$$

The iteration can be stopped, if $\Delta x_i \leq \delta \cdot x_i$. Since the expected statistical error for the HyperLogLog data structure scales according to $\frac{1}{\sqrt{m}}$ [6], it makes sense to choose $\delta = \frac{\varepsilon}{\sqrt{m}}$ with some constant ε . For all results that will be presented in Section 4.5 we have used $\varepsilon = 10^{-2}$.

4.4. Maximum likelihood estimation algorithm

In order to get a fast cardinality estimation algorithm, it is crucial to minimize the evaluation costs for (52). A couple of optimizations allow significant reduction of the computational effort:

- Only a fraction of all count values C_k is non-zero. If we denote $K_{\min} := \min\{k | C_k > 0\}$ and $K_{\max} := \max\{k | C_k > 0\}$, it is sufficient to loop over all indices in the range $[K_{\min}, K_{\max}]$.
- The sum $\sum_{k=0}^q \frac{C_k}{2^k}$ in (52) can be precalculated and reused for all function evaluations.
- Many programming languages allow the efficient multiplication and division by any integral power of two using special functions, such as `ldexp` in C/C++ or `scalb` in Java.
- The function $h(x)$ only needs to be evaluated at values $\left\{ \frac{x}{2^{K'_{\max}}}, \frac{x}{2^{K'_{\max}-1}}, \dots, \frac{x}{2^{K'_{\min}}} \right\}$ where $K'_{\max} := \min(K_{\max}, q)$. This series corresponds to a geometric series with ratio two. A straightforward calculation using (42) is very expensive because of the

exponential function. However, if we know $h\left(\frac{x}{2^{K'_{\max}}}\right)$ all other required function values can be easily obtained using the identity

$$h(4x) = \frac{x + h(2x)(1 - h(2x))}{x + (1 - h(2x))}. \quad (55)$$

This recursive formula is stable in a sense that the relative error of $h(4x)$ is smaller than that of $h(2x)$ as shown in Appendix A.

- If x is smaller than 0.5, the function $h(x)$ can be well approximated by a Taylor series around $x = 0$

$$h(x) = \frac{x}{2} - \frac{x^2}{12} + \frac{x^4}{720} - \frac{x^6}{30240} + \mathcal{O}(x^8), \quad (56)$$

which can be optimized for numerical evaluation using Estrin's scheme and $x' := \frac{x}{2}$ and $x'' := x'x'$

$$h(x) = x' - x''/3 + (x''x'') (1/45 - x''/472.5) + \mathcal{O}(x^8) \quad (57)$$

The smallest argument for which h needs to be evaluated is $\frac{x}{2^{K'_{\max}}}$. If $C_{q+1} = 0$, we can find an upper bound for the smallest argument using (48)

$$\frac{x}{2^{K'_{\max}}} \leq \log \left(1 + \frac{\sum_{k=0}^{K'_{\max}} C_k}{2^{K'_{\max}} \sum_{k=0}^{K'_{\max}} \frac{C_k}{2^k}} \right) \leq \log 2 \approx 0.693. \quad (58)$$

In practice, $\frac{x}{2^{K'_{\max}}} \leq 0.5$ is satisfied most of the time as long as only a few registers are saturated, that is $C_{q+1} \ll m$. In case $\frac{x}{2^{K'_{\max}}} > 0.5$, $h\left(\frac{x}{2^\kappa}\right)$ is calculated instead with $\kappa = 2 + \lfloor \log_2(x) \rfloor$. By definition, $\frac{x}{2^\kappa} \leq 0.5$ which allows using the Taylor series approximation. $h\left(\frac{x}{2^{K'_{\max}}}\right)$ is finally obtained after $\kappa - K'_{\max}$ iterations using (55). As shown in Appendix B, a small approximation error of h does not have much impact on the error of the maximum likelihood estimate as long as most registers are not yet saturated.

Putting all these optimizations together finally gives the new cardinality estimation algorithm as presented in Algorithm 4. The algorithm requires only elementary operations, except for large cardinalities, where it is worth to take the strong (45) instead of the weak lower bound (46) as initial value, which requires a single logarithm evaluation. The stronger bound is a much better approximation especially for large cardinalities and substantially reduces the number of required iteration cycles.

4.5. Estimation error

In order to investigate the estimation error for the maximum likelihood estimation algorithm, we investigated the same HyperLogLog configurations as in Section 3.5 for

Algorithm 4 Maximum likelihood cardinality estimation.

```

function ESTIMATECARDINALITY( $C$ )
   $q \leftarrow \dim(C) - 2$ 
   $K_{\min} \leftarrow \min\{k | C_k > 0\}$ 
  if  $K_{\min} > q$  then
    return  $\infty$ 
  end if
   $K'_{\min} \leftarrow \max(K_{\min}, 1)$ 
   $K_{\max} \leftarrow \max\{k | C_k > 0\}$ 
   $K'_{\max} \leftarrow \min(K_{\max}, q)$ 
   $z \leftarrow 0$ 
   $m' \leftarrow C_{q+1}$ 
   $y \leftarrow 2^{-K'_{\max}}$ 
  for  $k \leftarrow K'_{\max}, K'_{\min}$  do
     $z \leftarrow z + C_k \cdot y$   $\triangleright$  here  $y = 2^{-k}$ 
     $y \leftarrow 2y$ 
     $m' \leftarrow m' + C_k$ 
  end for
   $\triangleright$  here  $z = \sum_{k=1}^q \frac{C_k}{2^k}$ 
   $m \leftarrow m' + C_0$ 
   $c \leftarrow C_{q+1}$ 
  if  $q \geq 1$  then
     $c \leftarrow c + C_{K'_{\max}}$ 
  end if
   $g_{\text{prev}} \leftarrow 0$ 
   $a \leftarrow z + C_0$   $\triangleright a = \sum_{k=0}^q \frac{C_k}{2^k}$ 
   $b \leftarrow z + C_{q+1} \cdot 2^{-q}$   $\triangleright b = \sum_{k=1}^q \frac{C_k}{2^k} + \frac{C_{q+1}}{2^q}$ 
  if  $b \leq 1.5 \cdot a$  then
     $x \leftarrow m' / (0.5 \cdot b + a)$   $\triangleright$  weak lower bound (46)
  else
     $x \leftarrow m' / b \cdot \log(1 + b/a)$   $\triangleright$  strong lower bound (45)
  end if

```

Algorithm 4 Maximum likelihood cardinality estimation (continued).

```

 $\Delta x \leftarrow x$ 
 $\delta \leftarrow \varepsilon / \sqrt{m}$  ▷  $\varepsilon = 10^{-2}$  (see Section 4.3)
while  $\Delta x > x \cdot \delta$  do ▷ secant method iteration
     $\kappa \leftarrow 2 + \lfloor \log_2(x) \rfloor$ 
     $x' \leftarrow x \cdot 2^{-\max(K'_{\max}, \kappa) - 1}$  ▷  $x' \in [0, 0.25]$ 
     $x'' \leftarrow x' \cdot x'$ 
     $h \leftarrow x' - x''/3 + (x'' \cdot x'') \cdot (1/45 - x''/472.5)$  ▷ Taylor approximation (57)
    for  $k \leftarrow (\kappa - 1), K'_{\max}$  do
         $h \leftarrow \frac{x' + h \cdot (1-h)}{x' + (1-h)}$  ▷ calculate  $h(\frac{x}{2^k})$ , see (55), at  
this point  $x' = \frac{x}{2^{k+2}}$ 

         $x' \leftarrow 2x'$ 
    end for
     $g \leftarrow c \cdot h$  ▷ compare (52)
    for  $k \leftarrow (K'_{\max} - 1), K'_{\min}$  do
         $h \leftarrow \frac{x' + h \cdot (1-h)}{x' + (1-h)}$  ▷ calculate  $h(\frac{x}{2^k})$ , see (55), at  
this point  $x' = \frac{x}{2^{k+2}}$ 

         $g \leftarrow g + C_k \cdot h$ 
         $x' \leftarrow 2x'$ 
    end for
     $g \leftarrow g + x \cdot a$ 
    if  $g > g_{\text{prev}} \wedge m' \geq g$  then
         $\Delta x \leftarrow \Delta x \cdot \frac{m' - g}{g - g_{\text{prev}}}$  ▷ see (53)
    else
         $\Delta x \leftarrow 0$ 
    end if
     $x \leftarrow x + \Delta x$ 
     $g_{\text{prev}} \leftarrow g$ 
end while
return  $m \cdot x$ 
end function

```

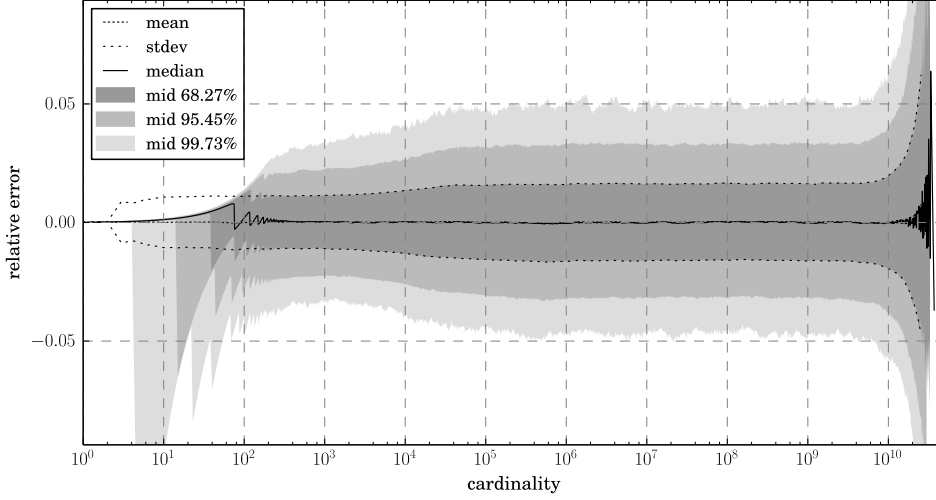


Figure 14: Relative error of the maximum likelihood estimates as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 12$ and $q = 20$.

the corrected raw estimation algorithm. Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, and Figure 19 show very similar results for various HyperLogLog parameters.

What is different for the maximum likelihood estimation approach is a somewhat smaller median bias with less oscillations around zero for small cardinalities. Furthermore, contrary to the raw estimator which reveals a small oscillating bias for the mean (see Figure 9), the maximum likelihood estimator seems to be completely unbiased (see Figure 17).

4.6. Performance

We also measured the performance of Algorithm 4 using the same test setup as described in Section 3.6. The results for HyperLogLog configurations $p = 12, q = 20$ and $p = 12, q = 52$ are shown in Figure 20. The average computation time for the maximum likelihood algorithm shows a different behavior than for the corrected raw estimation algorithm (compare Figure 12). As can be seen, the average execution time is larger for most cardinalities, but nevertheless, it never exceeds 700 ns which is still fast enough for many applications.

5. Cardinality estimation of set intersections and differences

While the union of two sets that are represented by HyperLogLog sketches can be straightforwardly computed (1), the computation of cardinalities of other set operations like intersections and differences is more challenging. A common approach to calculate

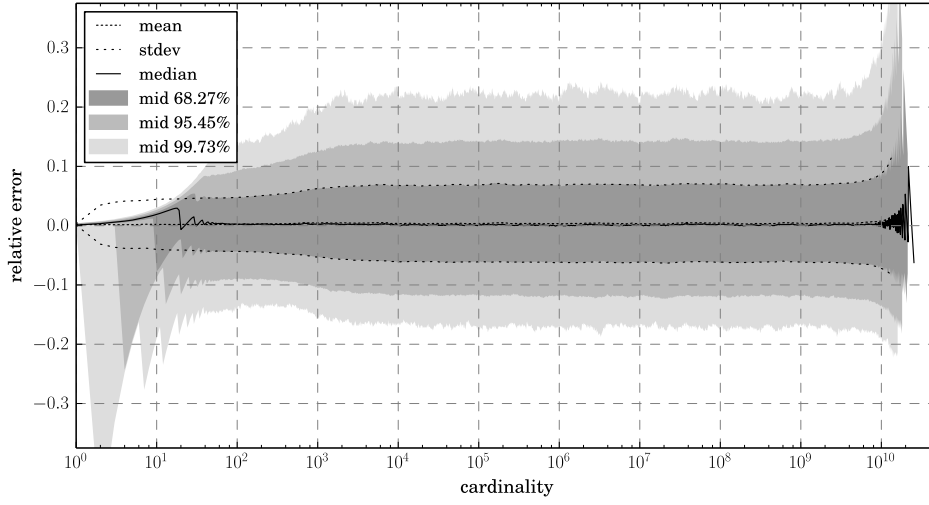


Figure 15: Relative error of the maximum likelihood estimates as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 8$ and $q = 24$.

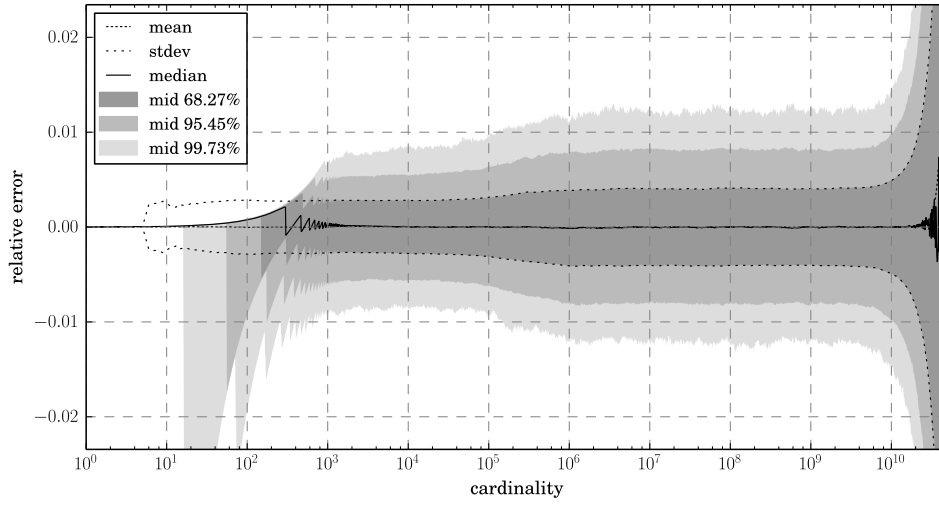


Figure 16: Relative error of the maximum likelihood estimates as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 16$ and $q = 16$.

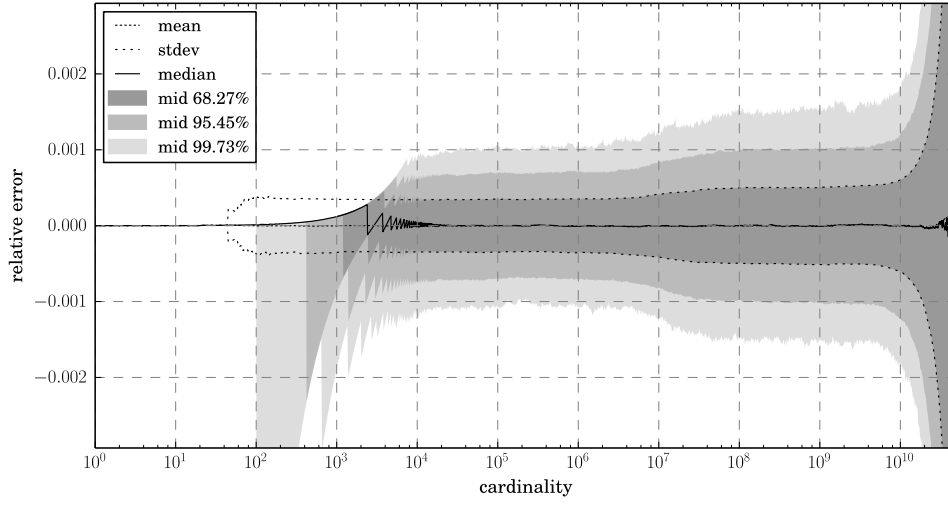


Figure 17: Relative error of the maximum likelihood estimates as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 22$ and $q = 10$.

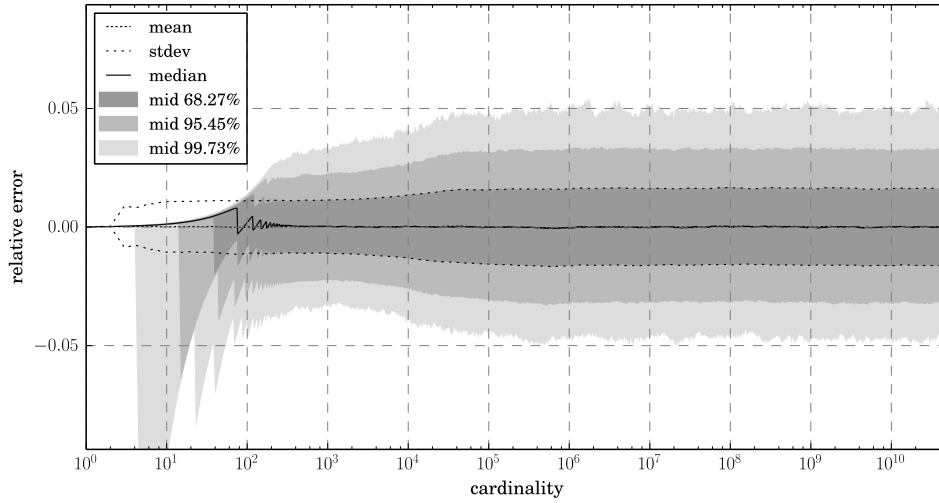


Figure 18: Relative error of the maximum likelihood estimates as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 12$ and $q = 52$.

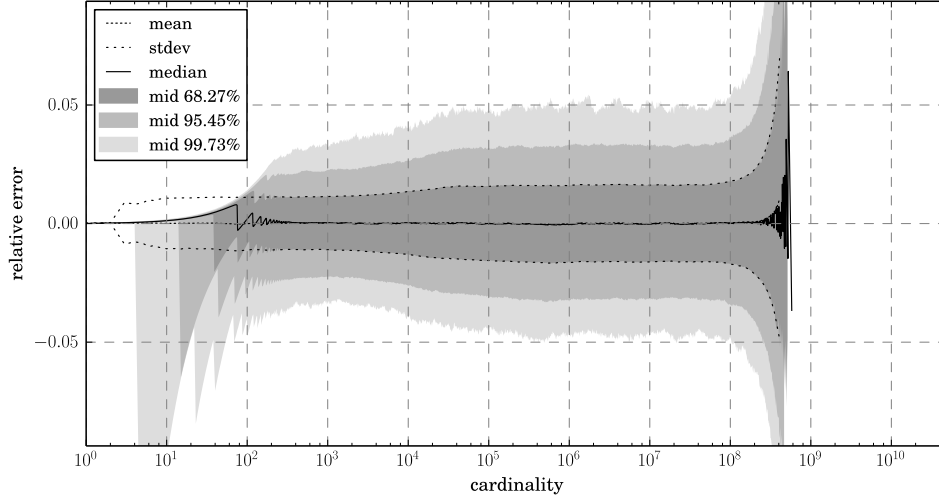


Figure 19: Relative error of the maximum likelihood estimates as a function of the true cardinality for a HyperLogLog sketch with parameters $p = 12$ and $q = 14$.

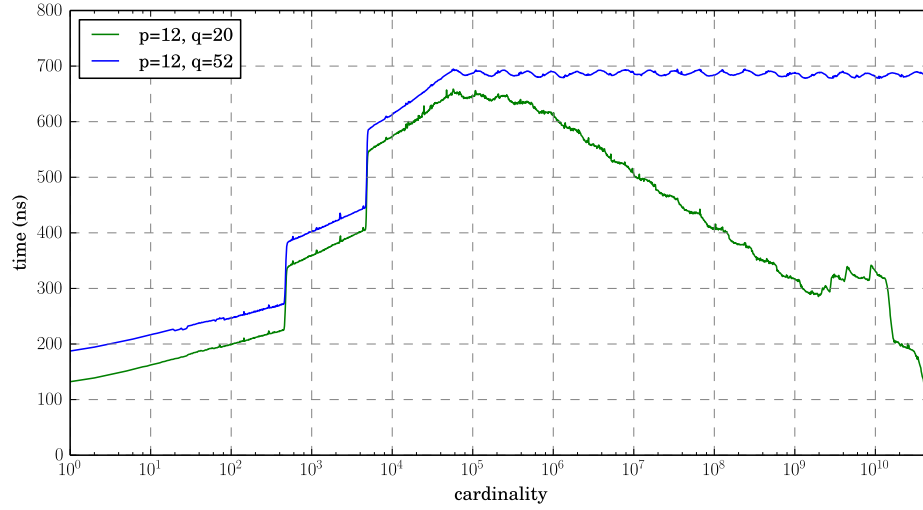


Figure 20: Average execution time of the maximum likelihood estimation algorithm as a function of the true cardinality with an Intel Core i5-2500K clocking at 3.3 GHz for HyperLogLog sketches with parameters $p = 12$, $q = 20$ and $p = 12$, $q = 52$, respectively.

the cardinality of intersections is the inclusion-exclusion principle

$$|S_1 \cap S_2| = |S_1| + |S_2| - |S_1 \cup S_2|. \quad (59)$$

Unfortunately, this approach can be very inaccurate. Due to estimation errors of HyperLogLog sketches the result of (59) can even be negative.

It was recently pointed out without special focus on HyperLogLog sketches, that the application of the maximum likelihood method to the joint likelihood function of two probabilistic data structures, which represent the intersection operands, gives better intersection size estimates [18]. For HyperLogLog sketches recorded without stochastic averaging (compare Section 2.1) this was also shown in [19], where also an algorithm based on the maximum likelihood principle was outlined. However, in practice, HyperLogLog sketches are recorded using stochastic averaging because of the much cheaper element insertions as described by Algorithm 1. Motivated by the good results we have obtained for a single HyperLogLog sketch using the maximum likelihood method in combination with the Poisson approximation, we are tempted to apply this approach also for the estimation of set operation result sizes.

Assume two given HyperLogLog sketches with register values \mathbf{K}_1 and \mathbf{K}_2 representing the sets S_1 and S_2 , respectively. The goal is to find estimates for the cardinalities of the pair-wise disjoint sets $X = S_1 \cap S_2$, $A = S_1 \setminus S_2$, and $B = S_2 \setminus S_1$. The Poisson approximation allows us to assume that the elements of A are inserted into the HyperLogLog sketch representing S_1 at rate λ_a . Similarly, the elements of B are inserted into the HyperLogLog sketch representing S_2 at rate λ_b . Furthermore, we assume that elements of X are inserted into both HyperLogLog sketches simultaneously at rate λ_x . We expect that good estimates $\hat{\lambda}_a$, $\hat{\lambda}_b$, and $\hat{\lambda}_x$ for the rates are also good estimates for the cardinalities of A , B , and X .

5.1. Joint log-likelihood function

In order to get maximum-likelihood estimators for $\hat{\lambda}_a$, $\hat{\lambda}_b$, and $\hat{\lambda}_x$ we need to derive the joint probability distribution of two HyperLogLog sketches. Under the Poisson model the register values are independent and identically distributed. Therefore, we first derive the joint probability distribution for a single register that has value K_1 in the first HyperLogLog sketch representing S_1 and value K_2 in the second HyperLogLog sketch representing S_2 .

The HyperLogLog sketch that represents S_1 can be thought to be constructed from two HyperLogLog sketches representing A and X and merging both using (1). Analogously, the HyperLogLog sketch for S_2 could have been obtained from sketches for B and X . Let K_a , K_b , and K_x be the value of the considered register in the HyperLogLog sketch representing A , B , and X , respectively. The corresponding values in sketches for S_1 and S_2 are given by

$$K_1 = \max(K_a, K_x), \quad K_2 = \max(K_b, K_x). \quad (60)$$

Their joint cumulative probability function is given as

$$\begin{aligned}
P(K_1 \leq k_1 \wedge K_2 \leq k_2) &= P(\max(K_a, K_x) \leq k_1 \wedge \max(K_b, K_x) \leq k_2) \\
&= P(K_a \leq k_1 \wedge K_b \leq k_2 \wedge K_x \leq \min(k_1, k_2)) \\
&= P(K_a \leq k_1) P(K_b \leq k_2) P(K_x \leq \min(k_1, k_2)). \tag{61}
\end{aligned}$$

Here the last transformation used the independence of K_a , K_b , and K_x , because by definition, the sets A , B , and X are disjoint. Furthermore, under the Poisson model K_a , K_b , and K_x obey (7). If we take into account that elements are added to A , B , and X at rates λ_x , λ_a , and λ_b , respectively, the probability that a certain register has a value less than or equal to k_1 in the first HyperLogLog sketch and simultaneously a value less than or equal to k_2 in the second one can be written as

$$P(K_1 \leq k_1 \wedge K_2 \leq k_2) = \begin{cases} 0 & k_1 < 0 \vee k_2 < 0 \\ e^{-\frac{\lambda_a}{m2^{k_1}} - \frac{\lambda_b}{m2^{k_2}} - \frac{\lambda_x}{m2^{\min(k_1, k_2)}}} & 0 \leq k_1 \leq q \wedge 0 \leq k_2 \leq q \\ e^{-\frac{\lambda_b + \lambda_x}{m2^{k_2}}} & 0 \leq k_2 \leq q < k_1 \\ e^{-\frac{\lambda_a + \lambda_x}{m2^{k_1}}} & 0 \leq k_1 \leq q < k_2 \\ 1 & q < k_1 \wedge q < k_2 \end{cases} \tag{62}$$

The joint probability mass function for both register values can be calculated using

$$\begin{aligned}
\rho(k_1, k_2) &= P(K_1 \leq k_1 \wedge K_2 \leq k_2) - P(K_1 \leq k_1 - 1 \wedge K_2 \leq k_2) \\
&\quad - P(K_1 \leq k_1 \wedge K_2 \leq k_2 - 1) + P(K_1 \leq k_1 - 1 \wedge K_2 \leq k_2 - 1), \tag{63}
\end{aligned}$$

which finally gives

$$\rho(k_1, k_2) = \begin{cases} e^{-\frac{\lambda_a + \lambda_x}{m} - \frac{\lambda_b}{m2^{k_2}}} \left(1 - e^{-\frac{\lambda_b}{m2^{k_2}}}\right) & 0 = k_1 < k_2 \leq q \\ e^{-\frac{\lambda_a + \lambda_x}{m}} \left(1 - e^{-\frac{\lambda_b}{m2^q}}\right) & 0 = k_1 < k_2 = q + 1 \\ e^{-\frac{\lambda_a + \lambda_x}{m2^{k_1}} - \frac{\lambda_b}{m2^{k_2}}} \left(1 - e^{-\frac{\lambda_a + \lambda_x}{m2^{k_1}}}\right) \left(1 - e^{-\frac{\lambda_b}{m2^{k_2}}}\right) & 1 \leq k_1 < k_2 \leq q \\ e^{-\frac{\lambda_a + \lambda_x}{m2^{k_1}} - \frac{\lambda_a + \lambda_x}{m2^{k_1}}} \left(1 - e^{-\frac{\lambda_a + \lambda_x}{m2^{k_1}}}\right) \left(1 - e^{-\frac{\lambda_b}{m2^q}}\right) & 1 \leq k_1 < k_2 = q + 1 \\ e^{-\frac{\lambda_b + \lambda_x}{m} - \frac{\lambda_a}{m2^{k_1}}} \left(1 - e^{-\frac{\lambda_a}{m2^{k_1}}}\right) & 0 = k_2 < k_1 \leq q \\ e^{-\frac{\lambda_b + \lambda_x}{m}} \left(1 - e^{-\frac{\lambda_a}{m2^q}}\right) & 0 = k_2 < k_1 = q + 1 \\ e^{-\frac{\lambda_b + \lambda_x}{m2^{k_2}} - \frac{\lambda_a}{m2^{k_1}}} \left(1 - e^{-\frac{\lambda_b + \lambda_x}{m2^{k_2}}}\right) \left(1 - e^{-\frac{\lambda_a}{m2^{k_1}}}\right) & 1 \leq k_2 < k_1 \leq q \\ e^{-\frac{\lambda_b + \lambda_x}{m2^{k_2}} - \frac{\lambda_b + \lambda_x}{m2^{k_2}}} \left(1 - e^{-\frac{\lambda_b + \lambda_x}{m2^{k_2}}}\right) \left(1 - e^{-\frac{\lambda_a}{m2^q}}\right) & 1 \leq k_2 < k_1 = q + 1 \\ e^{-\frac{\lambda_a + \lambda_b + \lambda_x}{m}} & 0 = k_1 = k_2 \\ e^{-\frac{\lambda_a + \lambda_b + \lambda_x}{m2^k} - \frac{\lambda_a + \lambda_b + \lambda_x}{m2^k} - e^{-\frac{\lambda_a + \lambda_b + \lambda_x}{m2^k}} + e^{-\frac{\lambda_a + \lambda_b + \lambda_x}{m2^k}}} & 1 \leq k_1 = k_2 = k \leq q \\ 1 - e^{-\frac{\lambda_a + \lambda_x}{m2^q}} - e^{-\frac{\lambda_b + \lambda_x}{m2^q}} + e^{-\frac{\lambda_a + \lambda_b + \lambda_x}{m2^q}} & k_1 = k_2 = q + 1 \end{cases} \quad (64)$$

The logarithm of the joint probability mass function can be written using Iverson bracket notation ($[\text{true}] := 1$, $[\text{false}] := 0$) as

$$\begin{aligned} \log(\rho(k_1, k_2)) = & \log\left(1 - e^{-\frac{\lambda_a + \lambda_x}{m2^{k_1}}}\right) [1 \leq k_1 < k_2] + \log\left(1 - e^{-\frac{\lambda_a}{m2^{\min(k_1, q)}}}\right) [k_2 < k_1] \\ & + \log\left(1 - e^{-\frac{\lambda_b + \lambda_x}{m2^{k_2}}}\right) [1 \leq k_2 < k_1] + \log\left(1 - e^{-\frac{\lambda_b}{m2^{\min(k_2, q)}}}\right) [k_1 < k_2] \\ & + \log\left(1 - e^{-\frac{\lambda_a + \lambda_x}{m2^{\min(k_1, q)}}} - e^{-\frac{\lambda_b + \lambda_x}{m2^{\min(k_1, q)}}} + e^{-\frac{\lambda_a + \lambda_b + \lambda_x}{m2^{\min(k_1, q)}}}\right) [1 \leq k_1 = k_2] \\ & - \frac{\lambda_a}{m2^{k_1}} [k_1 \leq q] - \frac{\lambda_b}{m2^{k_2}} [k_2 \leq q] - \frac{\lambda_x}{m2^{\min(k_1, k_2)}} [k_1 \leq q \vee k_2 \leq q]. \end{aligned} \quad (65)$$

Since the values for different registers are independent under the Poisson model, we are now able to write the joint probability mass function for all registers in both Hyper-LogLog sketches

$$\rho(\mathbf{k}_1, \mathbf{k}_2) = \prod_{i=1}^m \prod_{j=1}^m \rho(k_{1i}, k_{2j}). \quad (66)$$

In order to get the maximum likelihood estimates $\hat{\lambda}_a$, $\hat{\lambda}_b$, and $\hat{\lambda}_x$ we need to maximize

the log-likelihood function given by

$$\log \mathcal{L}(\lambda_a, \lambda_b, \lambda_x | \mathbf{K}_1, \mathbf{K}_2) = \sum_{i=1}^m \sum_{j=1}^m \log(\rho(K_{1i}, K_{2j})) \quad (67)$$

Insertion of (65) results in

$$\begin{aligned} \log \mathcal{L}(\lambda_a, \lambda_b, \lambda_x | \mathbf{K}_1, \mathbf{K}_2) = & \sum_{k=1}^q \log \left(1 - e^{-\frac{\lambda_a + \lambda_x}{m2^k}} \right) C_{1k}^< + \log \left(1 - e^{-\frac{\lambda_b + \lambda_x}{m2^k}} \right) C_{2k}^< \\ & + \sum_{k=1}^{q+1} \log \left(1 - e^{-\frac{\lambda_a}{m2^{\min(k,q)}}} \right) C_{1k}^> + \log \left(1 - e^{-\frac{\lambda_b}{m2^{\min(k,q)}}} \right) C_{2k}^> \\ & + \sum_{k=1}^{q+1} \log \left(1 - e^{-\frac{\lambda_a + \lambda_x}{m2^{\min(k,q)}}} - e^{-\frac{\lambda_b + \lambda_x}{m2^{\min(k,q)}}} + e^{-\frac{\lambda_a + \lambda_b + \lambda_x}{m2^{\min(k,q)}}} \right) C_k^= \\ & - \frac{\lambda_a}{m} \sum_{k=0}^q \frac{C_{1k}^< + C_k^= + C_{1k}^>}{2^k} - \frac{\lambda_b}{m} \sum_{k=0}^q \frac{C_{2k}^< + C_k^= + C_{2k}^>}{2^k} - \frac{\lambda_x}{m} \sum_{k=0}^q \frac{C_{1k}^< + C_k^= + C_{2k}^<}{2^k}, \end{aligned} \quad (68)$$

where the constants $C_{1k}^<$, $C_{1k}^>$, $C_{2k}^<$, $C_{2k}^>$, and $C_k^=$ are defined as

$$\begin{aligned} C_{1k}^< &:= |\{i | k = K_{1i} < K_{2i}\}|, \\ C_{1k}^> &:= |\{i | k = K_{1i} > K_{2i}\}|, \\ C_{2k}^< &:= |\{i | k = K_{2i} < K_{1i}\}|, \\ C_{2k}^> &:= |\{i | k = K_{2i} > K_{1i}\}|, \\ C_k^= &:= |\{i | k = K_{1i} = K_{2i}\}|. \end{aligned} \quad (69)$$

These $5m$ values are sufficient for estimating λ_a , λ_b , and λ_x . Actually, the set of these constants can be further reduced, because $C_{10}^> = C_{20}^> = C_{1,q+1}^< = C_{2,q+1}^< = 0$ always holds.

The log-likelihood function (68) does not always have a strict global maximum point. For example, if all register values of the first HyperLogLog sketch are larger than the corresponding values in the second HyperLogLog sketch, that is $C_{1k}^< = C_k^= = C_{2k}^> = 0$ for all k , the function can be rewritten as sum of two functions, one dependent on λ_a and the other dependent on $\lambda_b + \lambda_x$. The maximum is obtained, if $\lambda_a = \hat{\lambda}_1$ and $\lambda_b + \lambda_x = \hat{\lambda}_2$. Here $\hat{\lambda}_1$ and $\hat{\lambda}_2$ are the maximum likelihood cardinality estimates for the first and second HyperLogLog sketch, respectively. This means that the maximum-likelihood makes no clear statement about the intersection size in this case. The estimate for λ_x could be anything between 0 and $\hat{\lambda}_2$. For comparison, the inclusion-exclusion approach would give $\hat{\lambda}_2$ as estimate. This result is questionable, because there it is not evident in this case, that the sets S_1 and S_2 really have common elements.

Since (68) is the generalization of (37) for two HyperLogLog sketches, the log-likelihood function for a single HyperLogLog sketch can be obtained by considering special cases.

For example, assuming $\lambda_x = 0$, which means that both sketches represent disjoint sets, (68) can be splitted into the sum of two unary functions with parameters λ_a and λ_b , respectively. As expected, both functions correspond to (37). Another extreme case is that both sketches represent identical sets. In this case $\lambda_a = \lambda_b = 0$ and all the register values must be pair-wise identical, which finally yields (37) for λ_x .

The inclusion-exclusion method does not use all the available information given by the sufficient statistic (69), because the estimator is a function of the three vectors $(\mathbf{C}_1^< + \mathbf{C}^= + \mathbf{C}_1^>)$, $(\mathbf{C}_2^< + \mathbf{C}^= + \mathbf{C}_2^>)$, and $(\mathbf{C}_1^> + \mathbf{C}^= + \mathbf{C}_2^>)$. In contrast, the maximum-likelihood method incorporates each value of the sufficient statistic (69) individually.

5.2. Computation of the maximum likelihood estimates

The maximum likelihood estimates can be obtained by maximizing (68). Since the three parameters are all non-negative, this is a constrained optimization problem. In order to get rid of these constraints, we use the transformation $\lambda = me^\varphi$. This mapping has also the nice property that relative accuracy limits are translated into absolute ones, because $\Delta\varphi = \Delta\lambda/\lambda$. Many optimization algorithm implementations allow the definition of absolute limits rather than relative ones.

The transformed log-likelihood function can be written as

$$\begin{aligned}
f(\varphi_a, \varphi_b, \varphi_x) &:= \log \mathcal{L}(me^{\varphi_a}, me^{\varphi_b}, me^{\varphi_x} | \mathbf{K}_1, \mathbf{K}_2) = \\
&+ \sum_{k=1}^q C_{1k}^< \log(z_{xk} + y_{xk}z_{ak}) + C_{2k}^< \log(z_{xk} + y_{xk}z_{bk}) \\
&+ \sum_{k=1}^q C_{1k}^> \log(z_{ak}) + C_{2k}^> \log(z_{bk}) + C_k^= \log(z_{xk} + y_{xk}z_{ak}z_{bk}) \\
&+ C_{1,q+1}^> \log(z_{aq}) + C_{2,q+1}^> \log(z_{bq}) + C_{q+1}^= \log(z_{xq} + y_{xq}z_{aq}z_{bq}) \\
&- \sum_{k=0}^q (C_{1k}^< + C_k^= + C_{1k}^>) x_{ak} + (C_{2k}^< + C_k^= + C_{2k}^>) x_{bk} + (C_{1k}^< + C_k^= + C_{2k}^<) x_{xk}.
\end{aligned} \tag{70}$$

Here we introduced the following expressions for simplification:

$$x_{*k} := \frac{e^{\varphi_*}}{2^k}, \quad y_{*k} := e^{-x_{*k}}, \quad z_{*k} := 1 - y_{*k}. \tag{71}$$

Quasi-Newton methods are commonly used for finding the maximum of such multi-dimensional functions. They require the calculation of the gradient $\nabla f = \left(\frac{\partial f}{\partial \varphi_a}, \frac{\partial f}{\partial \varphi_b}, \frac{\partial f}{\partial \varphi_x} \right)$ which is, using

$$\frac{\partial x_{*k}}{\partial \varphi_*} = x_{*k}, \quad \frac{\partial y_{*k}}{\partial \varphi_*} = -x_{*k}y_{*k}, \quad \frac{\partial z_{*k}}{\partial \varphi_*} = x_{*k}y_{*k}, \tag{72}$$

given by

$$\begin{aligned}
\frac{\partial f}{\partial \varphi_a} &= \sum_{k=1}^q C_{1k}^< \frac{y_{xk} x_{ak} y_{ak}}{z_{xk} + y_{xk} z_{ak}} + C_k^= \frac{y_{xk} x_{ak} y_{ak} z_{bk}}{z_{xk} + y_{xk} z_{ak} z_{bk}} + C_{1k}^> \frac{x_{ak} y_{ak}}{z_{ak}} \\
&\quad + C_{q+1}^= \frac{y_{xq} x_{aq} y_{aq} z_{bq}}{z_{xq} + y_{xq} z_{aq} z_{bq}} + C_{1,q+1}^> \frac{x_{aq} y_{aq}}{z_{aq}} - \sum_{k=0}^q (C_{1k}^< + C_k^= + C_{1k}^>) x_{ak}, \\
\frac{\partial f}{\partial \varphi_b} &= \sum_{k=1}^q C_{2k}^< \frac{y_{xk} x_{bk} y_{bk}}{z_{xk} + y_{xk} z_{bk}} + C_k^= \frac{y_{xk} z_{ak} x_{bk} y_{bk}}{z_{xk} + y_{xk} z_{ak} z_{bk}} + C_{2k}^> \frac{x_{bk} y_{bk}}{z_{bk}} \\
&\quad + C_{q+1}^= \frac{y_{xq} z_{aq} x_{bq} y_{bq}}{z_{xq} + y_{xq} z_{aq} z_{bq}} + C_{2,q+1}^> \frac{x_{bq} y_{bq}}{z_{bq}} - \sum_{k=0}^q (C_{2k}^< + C_k^= + C_{2k}^>) x_{bk}, \\
\frac{\partial f}{\partial \varphi_x} &= \sum_{k=1}^q C_{1k}^< \frac{x_{xk} y_{xk} y_{ak}}{z_{xk} + y_{xk} z_{ak}} + C_k^= \frac{x_{xk} y_{xk} (y_{ak} + z_{ak} y_{bk})}{z_{xk} + y_{xk} z_{bk} z_{ak}} + C_{2k}^< \frac{x_{xk} y_{xk} y_{bk}}{z_{xk} + y_{xk} z_{bk}} \\
&\quad + C_{q+1}^= \frac{x_{xq} y_{xq} (y_{aq} + z_{aq} y_{bq})}{z_{xq} + y_{xq} z_{aq} z_{bq}} - \sum_{k=0}^q (C_{1k}^< + C_k^= + C_{2k}^<) x_{xk}.
\end{aligned} \tag{73}$$

The calculation of (70) and its derivatives requires some care when calculating y_{*k} and z_{*k} . Since x_{*k} is nonnegative, we have $y_{*k}, z_{*k} \in [0, 1]$. In order to reduce the numerical error of z_{*k} for small x_{*k} , it is essential to use the function $\text{expm1}(x) := e^x - 1$ that is available in most programming languages. If x_{*k} is smaller than $\log(2)$, we calculate y_{*k} and z_{*k} via $z_{*k} = -\text{expm1}(-x_{*k})$ and $y_{*k} = 1 - z_{*k}$ and not as defined in (71). In this way the numerical error of both, y_{*k} and z_{*k} , is minimized and still only a single exponential function needs to be evaluated.

Apart from that, the numerical evaluation of (70) is straightforward. The arguments of all logarithms are in range $[0, 1]$. The case that some argument vanishes, which would cause the logarithm to be negative infinite, does not occur in practice. Consider for example the logarithm evaluation associated with $C_{1k}^<$ which is only relevant if $C_{1k}^< > 0$. In this case however, we can be certain that the cardinality of $A \cup X$ is at least 1. Therefore, it is expected that at least one of the two maximum likelihood estimates $\hat{\lambda}_a$ and $\hat{\lambda}_x$ is at least in the order of 1. If the optimization algorithm starts with appropriate initial values, function evaluations for which $\min(\lambda_a, \lambda_x) \ll 1$ are not expected. Furthermore, the argument of the logarithm can be bounded by $z_{xk} + y_{xk} z_{ak} \geq \min(z_{xk}, z_{ak}) = 1 - e^{-\frac{\min(\lambda_a, \lambda_x)}{m2^k}} \approx \frac{\min(\lambda_a, \lambda_x)}{m2^k}$. The last approximation is valid, if $\min(\lambda_a, \lambda_x) \ll m2^k$, and shows that the argument of the logarithm is large enough to be accurately represented by double-precision floating-point numbers. Similar argumentation holds for all other logarithmic terms and also the divisions that appear in (73).

5.3. Results

We used the popular Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [20] to solve this optimization problem. We used the implementation provided by the GNU scientific

Algorithm 5 Joint cardinality estimation.

```

function ESTIMATECARDINALITIES(  $C_1^>, C_1^<, C_2^>, C_2^<, C^=$  )
   $m \leftarrow \|C_1^< + C^= + C_1^>\|_1$ 
   $\hat{\lambda}_{ax} \leftarrow \text{ESTIMATECARDINALITY}(C_1^< + C^= + C_1^>)$ 
   $\hat{\lambda}_{bx} \leftarrow \text{ESTIMATECARDINALITY}(C_2^< + C^= + C_2^>)$ 
  if  $C_{10}^< + C_0^= + C_{20}^< = m$  then
    return  $(\hat{\lambda}_{ax}, \hat{\lambda}_{bx}, 0)$   $\triangleright A$  and  $B$  are disjoint and  $X$  is empty in this case
  end if
   $\lambda_{abx} \leftarrow \text{ESTIMATECARDINALITY}(C_1^> + C^= + C_2^>)$ 
   $\varphi_a = \log(\max(1, \hat{\lambda}_{abx} - \hat{\lambda}_{bx})/m)$ 
   $\varphi_b = \log(\max(1, \hat{\lambda}_{abx} - \hat{\lambda}_{ax})/m)$ 
   $\varphi_x = \log(\max(1, \hat{\lambda}_{ax} + \hat{\lambda}_{bx} - \hat{\lambda}_{abx})/m)$ 
   $\delta \leftarrow \varepsilon/\sqrt{m}$   $\triangleright \varepsilon = 10^{-2}$ 
  repeat
     $\varphi'_a \leftarrow \varphi_a, \varphi'_b \leftarrow \varphi_b, \varphi'_x \leftarrow \varphi_x$ 
     $\varphi_a, \varphi_b, \varphi_x \leftarrow \text{OPTIMIZATIONSTEP}(\varphi_a, \varphi_b, \varphi_x; f, \nabla f)$ 
  until  $\max(|\varphi_a - \varphi'_a|, |\varphi_b - \varphi'_b|, |\varphi_x - \varphi'_x|) \leq \delta$ 
  return  $(me^{\varphi_a}, me^{\varphi_b}, me^{\varphi_x})$ 
end function

```

library [21].

TODO see results Table 1, Table 2, Table 3, Table 4

6. Conclusion and future work

TODO

A. Numerical stability of recursion formula for $h(x)$

In order to investigate the error propagation of a single recursion step using (55) we define $h_1 := h(2x)$ and $h_2 := h(4x)$. The recursion formula simplifies to

$$h_2 = \frac{x + h_1(1 - h_1)}{x + (1 - h_1)}. \quad (74)$$

If h_1 is approximated by $\tilde{h}_1 = h_1(1 + \varepsilon_1)$ with relative error ε_1 , the recursion formula will give an approximation for h_2

$$\tilde{h}_2 = \frac{x + \tilde{h}_1(1 - \tilde{h}_1)}{x + (1 - \tilde{h}_1)} \quad (75)$$

Table 1: List of true cardinalities of the pair-wise disjoint sets A , B , and X , for which maximum likelihood estimation from two HyperLogLog sketches with parameters $p = 20$ and $q = 44$ representing sets $A \cup X$ and $B \cup X$ was investigated.

#	$ A $	$ B $	$ X $	$\frac{ X }{ A \cup B \cup X }$	$\log_{10}\left(\frac{ A \cup X }{ B \cup X }\right)$
1	1 598 728 349	251 188 153	2 742 179	1.480e-3	0.804
2	13 049 219 720	5 945 604 476	103 608 860	5.425e-3	0.341
3	144 199	1608	3457	2.316e-2	1.953
4	36 085 381	8 525 744	3 516 662	7.307e-2	0.627
5	652 906 563	42 736 075	23 996 891	3.335e-2	1.184
6	74 772	2617	235	3.027e-3	1.456
7	148 568	12 597	7147	4.246e-2	1.072
8	3 837 563 445	277 467 992	104 644 949	2.480e-2	1.141
9	84 256	1624	402	4.659e-3	1.715
10	22 160 727	6 987 227	586 521	1.973e-2	0.501
11	48 156 078	4 333 904	1 313 150	2.441e-2	1.046
12	184 928	117 006	27 370	8.311e-2	0.199
13	17 149	2540	1530	7.211e-2	0.829
14	19 279 012	2 939 988	873 249	3.782e-2	0.817
15	1 602 292	23 812	2277	1.398e-3	1.828
16	99 786	1923	2805	2.684e-2	1.715
17	61 756 895	744 726	974 256	1.535e-2	1.919
18	134 200 029	11 958 008	15 335 333	9.496e-2	1.050
19	110 226	1205	691	6.163e-3	1.961
20	773 240 139	13 474 584	20 465 059	2.535e-2	1.759
21	307 179	5046	8549	2.665e-2	1.784
22	74 032	16 317	6343	6.560e-2	0.657
23	50 221	4898	958	1.708e-2	1.011
24	1 097 816	452 822	48 262	3.018e-2	0.385
25	84 256	2919	2210	2.472e-2	1.460
26	13 049 219 720	721 215 037	145 319 400	1.044e-2	1.258
27	59 347 163	1 287 276	286 512	4.703e-3	1.664
28	119 095 712	5 235 826	586 521	4.695e-3	1.357
29	32 092	4054	272	7.469e-3	0.899
30	47 102 261 762	2 264 762 985	1 174 389 309	2.324e-2	1.318
31	15 837	13 915	1441	4.620e-2	0.056
32	3 083 083 913	1 174 389 309	162 128 255	3.668e-2	0.419
33	464	305	14	1.788e-2	0.182
34	3857	3224	87	1.214e-2	0.078
35	5 329 186 306	4 114 387 435	429 886 036	4.354e-2	0.112
36	2 138 265	34 751	6343	2.910e-3	1.789
37	9 703 090	3 733 007	18 023	1.340e-3	0.415
38	721 215 037	220 709 637	6 325 446	6.671e-3	0.514
39	26 507 497	8 275 003	1 479 690	4.081e-2	0.506
40	35 805	1707	146	3.877e-3	1.322
41	750 499 260	447 341 132	3 089 958	2.573e-3	0.225
42	8 441 331	1 634 498	53 310	5.263e-3	0.713
43	5 235 826	485 486	64 405	1.113e-2	1.033
44	93 072	52 782	358	2.449e-3	0.246
45	5 172 455 725	3 208 269 483	72 414 817	8.567e-3	0.207
46	1 697 082 357	131 555 758	4 509 877	2.460e-3	1.111

Table 2: The mean and the standard deviation for the cases given in Table 1 when estimating $|A|$ using the inclusion-exclusion principle and the maximum likelihood method, respectively.

#	inclusion-exclusion principle			maximum likelihood method			improvement
	mean	stdev	rmse	mean	stdev	rmse	rmse ratio
1	1.289e-5	1.164e-3	1.164e-3	9.253e-6	1.037e-3	1.037e-3	1.122
2	-2.285e-5	1.417e-3	1.417e-3	-6.259e-6	1.086e-3	1.086e-3	1.304
3	-7.719e-6	7.118e-4	7.118e-4	-7.999e-6	7.060e-4	7.061e-4	1.008
4	-6.395e-5	1.319e-3	1.320e-3	-4.976e-5	1.107e-3	1.108e-3	1.191
5	1.925e-5	1.101e-3	1.101e-3	2.873e-5	1.044e-3	1.044e-3	1.055
6	-1.374e-5	7.271e-4	7.272e-4	-1.286e-5	7.169e-4	7.171e-4	1.014
7	-1.073e-5	7.741e-4	7.742e-4	-8.754e-6	7.339e-4	7.340e-4	1.055
8	3.529e-5	1.106e-3	1.107e-3	3.125e-5	1.043e-3	1.043e-3	1.061
9	1.297e-5	7.218e-4	7.220e-4	1.100e-5	7.153e-4	7.154e-4	1.009
10	-5.030e-5	1.272e-3	1.273e-3	-4.358e-5	1.033e-3	1.034e-3	1.231
11	5.651e-6	1.127e-3	1.127e-3	6.768e-6	1.028e-3	1.028e-3	1.097
12	-1.255e-5	1.169e-3	1.169e-3	-2.489e-6	9.042e-4	9.042e-4	1.293
13	-8.608e-7	8.447e-4	8.447e-4	-1.320e-5	7.684e-4	7.685e-4	1.099
14	-3.134e-5	1.153e-3	1.153e-3	-2.060e-5	1.013e-3	1.013e-3	1.138
15	-1.300e-5	7.731e-4	7.732e-4	-1.086e-5	7.646e-4	7.646e-4	1.011
16	-1.344e-5	7.384e-4	7.385e-4	-1.339e-5	7.229e-4	7.230e-4	1.021
17	7.711e-6	1.022e-3	1.022e-3	1.166e-5	1.007e-3	1.008e-3	1.014
18	1.899e-5	1.241e-3	1.242e-3	2.097e-5	1.141e-3	1.141e-3	1.088
19	-4.536e-6	7.130e-4	7.130e-4	-3.445e-6	7.058e-4	7.058e-4	1.010
20	5.285e-6	1.058e-3	1.058e-3	1.027e-5	1.036e-3	1.036e-3	1.022
21	7.841e-6	7.298e-4	7.299e-4	8.424e-6	7.127e-4	7.128e-4	1.024
22	-1.129e-5	8.831e-4	8.832e-4	-4.706e-6	7.851e-4	7.851e-4	1.125
23	-1.347e-5	7.880e-4	7.881e-4	-8.253e-6	7.453e-4	7.453e-4	1.057
24	1.349e-5	1.079e-3	1.079e-3	1.914e-5	8.349e-4	8.351e-4	1.292
25	-1.781e-5	7.610e-4	7.613e-4	-1.347e-5	7.361e-4	7.362e-4	1.034
26	7.786e-6	1.085e-3	1.085e-3	2.385e-6	1.054e-3	1.054e-3	1.029
27	6.713e-6	1.017e-3	1.017e-3	1.233e-5	9.969e-4	9.970e-4	1.020
28	1.063e-5	1.090e-3	1.091e-3	1.499e-5	1.053e-3	1.053e-3	1.036
29	1.519e-5	7.680e-4	7.681e-4	1.337e-5	7.205e-4	7.206e-4	1.066
30	-1.724e-5	1.082e-3	1.082e-3	-2.117e-5	1.038e-3	1.038e-3	1.043
31	-2.077e-5	1.193e-3	1.193e-3	-1.564e-5	8.946e-4	8.947e-4	1.333
32	-9.366e-7	1.385e-3	1.385e-3	1.556e-5	1.112e-3	1.112e-3	1.246
33	-1.520e-5	1.078e-3	1.078e-3	-2.181e-5	8.660e-4	8.663e-4	1.244
34	1.625e-5	1.130e-3	1.131e-3	6.216e-6	8.676e-4	8.676e-4	1.303
35	-3.313e-6	1.665e-3	1.665e-3	1.690e-5	1.181e-3	1.181e-3	1.409
36	-1.286e-5	8.168e-4	8.170e-4	-1.227e-5	8.034e-4	8.035e-4	1.017
37	-1.348e-5	1.300e-3	1.300e-3	4.443e-5	1.014e-3	1.015e-3	1.280
38	5.351e-5	1.283e-3	1.284e-3	2.409e-5	1.048e-3	1.048e-3	1.225
39	-7.546e-5	1.332e-3	1.335e-3	-5.088e-5	1.074e-3	1.075e-3	1.241
40	-4.855e-6	7.332e-4	7.332e-4	-4.539e-6	7.145e-4	7.145e-4	1.026
41	7.932e-6	1.475e-3	1.475e-3	1.590e-5	1.062e-3	1.062e-3	1.389
42	-1.825e-5	1.144e-3	1.144e-3	1.521e-6	9.694e-4	9.694e-4	1.180
43	-2.281e-6	9.927e-4	9.927e-4	5.092e-6	9.254e-4	9.254e-4	1.073
44	1.395e-5	1.034e-3	1.034e-3	-9.648e-6	7.984e-4	7.985e-4	1.296
45	-9.423e-6	1.529e-3	1.529e-3	4.252e-6	1.098e-3	1.098e-3	1.392
46	1.672e-5	1.094e-3	1.094e-3	8.882e-6	1.016e-3	1.016e-3	1.077

Table 3: The mean and the standard deviation for the cases given in Table 1 when estimating $|B|$ using the inclusion-exclusion principle and the maximum likelihood method, respectively.

#	inclusion-exclusion principle			maximum likelihood method			improvement
	mean	stdev	rmse	mean	stdev	rmse	rmse ratio
1	3.035e-5	3.698e-3	3.698e-3	6.687e-6	1.580e-3	1.580e-3	2.340
2	-4.702e-5	2.302e-3	2.302e-3	-1.229e-5	1.215e-3	1.215e-3	1.895
3	-4.868e-5	9.643e-3	9.643e-3	-7.699e-6	7.152e-3	7.152e-3	1.348
4	-9.060e-5	3.298e-3	3.299e-3	-4.662e-5	2.022e-3	2.023e-3	1.631
5	-1.553e-4	5.737e-3	5.739e-3	-3.377e-5	3.482e-3	3.482e-3	1.648
6	-5.567e-5	5.240e-3	5.240e-3	-2.696e-5	3.184e-3	3.184e-3	1.646
7	-3.752e-5	3.541e-3	3.541e-3	-1.710e-5	2.453e-3	2.453e-3	1.444
8	8.603e-5	5.481e-3	5.482e-3	4.682e-5	3.134e-3	3.134e-3	1.749
9	1.340e-4	7.143e-3	7.144e-3	3.964e-5	4.489e-3	4.489e-3	1.591
10	-5.585e-5	2.734e-3	2.734e-3	-3.715e-5	1.453e-3	1.454e-3	1.881
11	2.653e-5	5.045e-3	5.045e-3	5.603e-6	2.734e-3	2.734e-3	1.846
12	2.204e-5	1.546e-3	1.546e-3	3.769e-5	1.110e-3	1.111e-3	1.392
13	-1.820e-5	2.754e-3	2.754e-3	-7.230e-5	1.920e-3	1.921e-3	1.434
14	-3.790e-5	3.875e-3	3.875e-3	2.126e-5	2.185e-3	2.185e-3	1.773
15	-2.034e-4	9.797e-3	9.799e-3	-6.541e-5	4.727e-3	4.728e-3	2.073
16	-1.441e-4	7.286e-3	7.287e-3	-1.365e-4	5.392e-3	5.394e-3	1.351
17	1.020e-4	1.335e-2	1.335e-2	1.307e-4	8.990e-3	8.991e-3	1.485
18	-5.421e-5	5.135e-3	5.135e-3	-1.322e-5	3.587e-3	3.587e-3	1.432
19	2.219e-5	9.444e-3	9.444e-3	7.060e-5	6.522e-3	6.522e-3	1.448
20	9.566e-5	1.069e-2	1.069e-2	1.670e-4	7.528e-3	7.530e-3	1.419
21	-1.068e-4	8.116e-3	8.117e-3	-7.052e-5	5.904e-3	5.905e-3	1.375
22	-2.241e-5	2.316e-3	2.316e-3	1.208e-6	1.617e-3	1.617e-3	1.433
23	-3.027e-5	3.229e-3	3.229e-3	1.439e-5	2.058e-3	2.058e-3	1.569
24	4.772e-5	1.982e-3	1.982e-3	6.275e-5	1.200e-3	1.202e-3	1.649
25	-2.276e-4	5.571e-3	5.575e-3	-1.193e-4	3.904e-3	3.906e-3	1.427
26	3.200e-5	6.155e-3	6.155e-3	-4.054e-5	3.202e-3	3.203e-3	1.922
27	-2.787e-4	9.641e-3	9.645e-3	-7.746e-5	4.875e-3	4.875e-3	1.978
28	-1.264e-4	7.032e-3	7.033e-3	-3.781e-5	3.128e-3	3.128e-3	2.248
29	-3.365e-5	2.864e-3	2.864e-3	-4.790e-5	1.809e-3	1.810e-3	1.583
30	1.726e-5	6.839e-3	6.839e-3	-3.453e-5	3.976e-3	3.976e-3	1.720
31	-3.148e-6	1.305e-3	1.305e-3	3.896e-6	9.825e-4	9.825e-4	1.328
32	-4.329e-5	2.538e-3	2.538e-3	-3.585e-6	1.502e-3	1.502e-3	1.690
33	-5.696e-6	1.388e-3	1.388e-3	-1.522e-5	9.998e-4	9.999e-4	1.389
34	1.558e-5	1.273e-3	1.273e-3	3.380e-6	9.479e-4	9.479e-4	1.343
35	2.107e-5	1.959e-3	1.959e-3	4.770e-5	1.281e-3	1.282e-3	1.528
36	-4.961e-5	9.654e-3	9.654e-3	-3.448e-5	5.003e-3	5.003e-3	1.930
37	-3.169e-5	2.480e-3	2.480e-3	1.185e-4	1.328e-3	1.333e-3	1.860
38	1.128e-4	2.823e-3	2.825e-3	1.680e-5	1.365e-3	1.365e-3	2.069
39	-9.187e-5	2.873e-3	2.875e-3	-2.070e-5	1.628e-3	1.628e-3	1.766
40	2.855e-6	4.558e-3	4.558e-3	1.170e-5	2.769e-3	2.769e-3	1.646
41	-3.915e-5	2.128e-3	2.128e-3	-2.596e-5	1.167e-3	1.168e-3	1.823
42	-6.558e-5	3.329e-3	3.330e-3	3.486e-5	1.408e-3	1.409e-3	2.364
43	-6.749e-5	4.581e-3	4.581e-3	9.777e-6	2.176e-3	2.176e-3	2.106
44	1.981e-5	1.507e-3	1.507e-3	-2.162e-5	9.530e-4	9.533e-4	1.581
45	-1.831e-5	2.086e-3	2.086e-3	3.148e-6	1.204e-3	1.204e-3	1.733
46	9.727e-5	5.202e-3	5.203e-3	-4.435e-7	2.079e-3	2.079e-3	2.503

Table 4: The mean and the standard deviation for the cases given in Table 1 when estimating $|X|$ using the inclusion-exclusion principle and the maximum likelihood method, respectively.

#	inclusion-exclusion principle			maximum likelihood method			improvement
	mean	stdev	rmse	mean	stdev	rmse	rmse ratio
1	-2.437e-3	3.264e-1	3.264e-1	-3.720e-4	1.121e-1	1.121e-1	2.912
2	3.662e-3	1.197e-1	1.198e-1	1.688e-3	4.011e-2	4.015e-2	2.983
3	4.081e-5	4.504e-3	4.504e-3	2.090e-5	3.370e-3	3.370e-3	1.337
4	1.653e-4	7.648e-3	7.649e-3	5.769e-5	4.399e-3	4.400e-3	1.739
5	2.478e-4	1.009e-2	1.009e-2	3.039e-5	6.060e-3	6.060e-3	1.666
6	3.974e-4	5.767e-2	5.767e-2	2.027e-4	3.455e-2	3.455e-2	1.669
7	7.918e-5	5.995e-3	5.996e-3	3.872e-5	4.110e-3	4.110e-3	1.459
8	-1.996e-4	1.427e-2	1.428e-2	-9.579e-5	7.874e-3	7.875e-3	1.813
9	-5.706e-4	2.869e-2	2.869e-2	-1.432e-4	1.788e-2	1.788e-2	1.604
10	3.700e-4	3.070e-2	3.070e-2	1.506e-4	1.357e-2	1.357e-2	2.262
11	-1.204e-4	1.631e-2	1.631e-2	-5.351e-5	8.538e-3	8.538e-3	1.910
12	-5.165e-5	5.661e-3	5.662e-3	-1.142e-4	3.608e-3	3.610e-3	1.568
13	4.061e-6	4.318e-3	4.318e-3	9.119e-5	2.998e-3	2.999e-3	1.440
14	1.723e-5	1.267e-2	1.267e-2	-1.815e-4	6.876e-3	6.879e-3	1.841
15	2.057e-3	1.020e-1	1.020e-1	6.241e-4	4.868e-2	4.868e-2	2.096
16	1.115e-4	4.986e-3	4.987e-3	1.047e-4	3.727e-3	3.728e-3	1.338
17	-9.800e-5	1.022e-2	1.022e-2	-1.223e-4	6.881e-3	6.882e-3	1.485
18	3.215e-5	4.010e-3	4.010e-3	7.973e-7	2.844e-3	2.844e-3	1.410
19	-7.916e-5	1.641e-2	1.641e-2	-2.091e-4	1.130e-2	1.130e-2	1.453
20	-1.385e-5	7.100e-3	7.100e-3	-6.244e-5	5.032e-3	5.032e-3	1.411
21	7.833e-5	4.726e-3	4.726e-3	5.595e-5	3.479e-3	3.480e-3	1.358
22	4.137e-5	5.476e-3	5.476e-3	-1.908e-5	3.675e-3	3.676e-3	1.490
23	1.634e-4	1.615e-2	1.615e-2	-1.080e-4	9.955e-3	9.955e-3	1.622
24	-5.498e-4	1.718e-2	1.719e-2	-6.868e-4	8.998e-3	9.024e-3	1.905
25	2.729e-4	7.261e-3	7.266e-3	1.323e-4	5.069e-3	5.071e-3	1.433
26	-1.751e-4	3.005e-2	3.005e-2	1.841e-4	1.501e-2	1.501e-2	2.002
27	1.183e-3	4.327e-2	4.329e-2	2.797e-4	2.165e-2	2.165e-2	2.000
28	1.164e-3	6.230e-2	6.231e-2	3.712e-4	2.689e-2	2.689e-2	2.317
29	1.550e-4	4.123e-2	4.123e-2	3.329e-4	2.466e-2	2.466e-2	1.672
30	-3.329e-5	1.308e-2	1.308e-2	6.740e-5	7.447e-3	7.447e-3	1.756
31	-3.854e-6	1.020e-2	1.020e-2	-6.418e-5	6.265e-3	6.266e-3	1.629
32	3.896e-4	1.700e-2	1.701e-2	1.019e-4	8.108e-3	8.109e-3	2.097
33	7.621e-5	2.619e-2	2.619e-2	2.808e-4	1.552e-2	1.553e-2	1.687
34	-7.173e-4	3.911e-2	3.912e-2	-2.724e-4	2.339e-2	2.339e-2	1.672
35	1.411e-4	1.590e-2	1.590e-2	-1.060e-4	7.543e-3	7.544e-3	2.108
36	3.007e-4	5.277e-2	5.277e-2	2.351e-4	2.711e-2	2.711e-2	1.947
37	1.080e-2	4.686e-1	4.688e-1	-2.335e-2	2.055e-1	2.068e-1	2.267
38	-4.057e-3	9.101e-2	9.110e-2	-6.975e-4	3.193e-2	3.194e-2	2.852
39	3.479e-4	1.503e-2	1.503e-2	-4.898e-5	7.428e-3	7.429e-3	2.024
40	-2.737e-5	5.257e-2	5.257e-2	-4.186e-5	3.124e-2	3.124e-2	1.683
41	5.752e-3	2.671e-1	2.672e-1	3.879e-3	8.066e-2	8.075e-2	3.309
42	2.003e-3	9.923e-2	9.925e-2	-1.075e-3	3.601e-2	3.603e-2	2.755
43	3.401e-4	3.392e-2	3.393e-2	-2.392e-4	1.543e-2	1.543e-2	2.198
44	-3.934e-3	1.954e-1	1.955e-1	2.126e-3	9.518e-2	9.520e-2	2.053
45	1.753e-3	8.153e-2	8.155e-2	8.132e-4	2.944e-2	2.945e-2	2.769
46	-2.605e-3	1.488e-1	1.489e-1	2.447e-4	5.350e-2	5.350e-2	2.782

The corresponding relative error ε_2 is given by

$$\varepsilon_2 = \frac{\tilde{h}_2}{h_2} - 1. \quad (76)$$

Combination of (74), (75), and (76) yields for its absolute value

$$|\varepsilon_2| = |\varepsilon_1| \frac{\left| \frac{h_1(1-2h_1)}{x+h_1(1-h_1)} + \frac{h_1}{x+1-h_1} - \varepsilon_1 \frac{h_1^2}{x+h_1(1-h_1)} \right|}{\left| 1 - \varepsilon_1 \frac{h_1}{x+1-h_1} \right|}. \quad (77)$$

The triangle inequality leads to

$$|\varepsilon_2| \leq |\varepsilon_1| \frac{\left| \frac{h_1(1-2h_1)}{x+h_1(1-h_1)} + \frac{h_1}{x+1-h_1} \right| + |\varepsilon_1| \frac{h_1^2}{x+h_1(1-h_1)}}{\left| 1 - \varepsilon_1 \frac{h_1}{x+1-h_1} \right|} \quad (78)$$

By numerical means it is easy to show that the inequalities $\left| \frac{h_1(1-2h_1)}{x+h_1(1-h_1)} + \frac{h_1}{x+1-h_1} \right| \leq 0.517$, $\frac{h_1^2}{x+h_1(1-h_1)} \leq 0.436$, and $\frac{h_1}{x+1-h_1} \leq 0.529$ hold for all $x \geq 0$. Therefore, if we additionally assume, for example, $|\varepsilon_1| \leq 0.1$, we get

$$|\varepsilon_2| \leq |\varepsilon_1| \frac{0.517 + 0.1 \cdot 0.436}{1 - 0.1 \cdot 0.529} \leq |\varepsilon_1| \cdot 0.592, \quad (79)$$

which means that the relative error is decreasing in each recursion step and the recursive calculation of h is numerically stable.

B. Error caused by approximation of $h(x)$

According to (41) the exact estimate \hat{x} fulfills

$$\hat{x} \sum_{k=0}^q \frac{C_k}{2^k} + \sum_{k=1}^q C_k h \left(\frac{\hat{x}}{2^k} \right) + C_{q+1} h \left(\frac{\hat{x}}{2^q} \right) = m - C_0 \quad (80)$$

If h is not calculated exactly but approximated by \tilde{h} with maximum relative error $\varepsilon_h \ll 1$

$$\left| \tilde{h}(x) - h(x) \right| \leq \varepsilon_h h(x) \quad (81)$$

the solution of the equation will be off by a relative error ε_x :

$$\hat{x} (1 + \varepsilon_x) \sum_{k=0}^q \frac{C_k}{2^k} + \sum_{k=1}^q C_k \tilde{h} \left(\frac{\hat{x} (1 + \varepsilon_x)}{2^k} \right) + C_{q+1} \tilde{h} \left(\frac{\hat{x} (1 + \varepsilon_x)}{2^q} \right) = m - C_0 \quad (82)$$

Due to (81) there exists some $\alpha \in [-\varepsilon_h, \varepsilon_h]$ for which

$$\hat{x} (1 + \varepsilon_x) \sum_{k=0}^q \frac{C_k}{2^k} + (1 + \alpha) \left(\sum_{k=1}^q C_k h \left(\frac{\hat{x} (1 + \varepsilon_x)}{2^k} \right) + C_{q+1} h \left(\frac{\hat{x} (1 + \varepsilon_x)}{2^q} \right) \right) = m - C_0 \quad (83)$$

Since $h'(x) \in [0, 0.5]$ for $x \geq 0$, there exists a $\beta \in [0, 0.5]$ for which

$$\begin{aligned} \hat{x}(1 + \varepsilon_x) \sum_{k=0}^q \frac{C_k}{2^k} + (1 + \alpha) \left(\sum_{k=1}^q C_k h\left(\frac{\hat{x}}{2^k}\right) + \frac{C_k}{2^k} \hat{x} \varepsilon_x \beta \right) + \\ + (1 + \alpha) \left(C_{q+1} h\left(\frac{\hat{x}}{2^q}\right) + \frac{C_{q+1}}{2^q} \hat{x} \varepsilon_x \beta \right) = m - C_0 \end{aligned} \quad (84)$$

Subtracting (80) multiplied by $(1 + \alpha)$ from (84) and resolving ε_x gives

$$\varepsilon_x = \alpha \frac{\hat{x} \sum_{k=0}^q \frac{C_k}{2^k} - (m - C_0)}{\hat{x} \left(\sum_{k=0}^q \frac{C_k}{2^k} + (1 + \alpha) \beta \left(\sum_{k=1}^q \frac{C_k}{2^k} + \frac{C_{q+1}}{2^q} \right) \right)} \quad (85)$$

Using $|\alpha| \leq \varepsilon_h$, $\beta \geq 0$, and (49) the absolute value of the relative error can be bounded by

$$|\varepsilon_x| \leq |\varepsilon_h| \frac{(m - C_0) - \hat{x} \sum_{k=0}^q \frac{C_k}{2^k}}{\hat{x} \sum_{k=0}^q \frac{C_k}{2^k}} \quad (86)$$

Furthermore, using (46) we finally get

$$|\varepsilon_x| \leq |\varepsilon_h| \frac{\frac{1}{2} \sum_{k=1}^q \frac{C_k}{2^k} + \frac{C_{q+1}}{2^q}}{\sum_{k=0}^q \frac{C_k}{2^k}} \leq |\varepsilon_h| \left(\frac{1}{2} + \frac{\frac{C_{q+1}}{2^q}}{\sum_{k=0}^q \frac{C_k}{2^k}} \right) \leq |\varepsilon_h| \left(\frac{1}{2} + \frac{C_{q+1}}{m - C_{q+1}} \right) \quad (87)$$

Hence, as long as most registers are not in the saturated state ($C_{q+1} \ll m$), the relative error ε_x of the calculated estimate using the approximation $\tilde{h}(x)$ for $h(x)$ has the same order of magnitude as ε_h .

References

- [1] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.
- [2] Ahmed Metwally, Divyakant Agrawal, and Amr El Abbadi. Why go logarithmic if we can go linear?: Towards effective distinct counting of search traffic. In *Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology*, pages 618–629, Nantes, France, March 2008.
- [3] Daniel Ting. Streamed approximate counting of distinct elements: Beating optimal batch methods. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 442–451, New York, NY, USA, August 2014.
- [4] Piotr Indyk and David Woodruff. Tight lower bounds for the distinct elements problem. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 283–288, Cambridge, MA, USA, October 2003.

- [5] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 41–52, Indianapolis, IN, USA, June 2010.
- [6] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm. In *Proceedings of the 13th Conference on Analysis of Algorithms*, pages 127–146, Juan des Pins, France, June 2007.
- [7] Stefan Heule, Marc Nunkesser, and Alexander Hall. Hyperloglog in practice: Algorithmic engineering of a state of the art cardinality estimation algorithm. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 683–692, Genoa, Italy, March 2013.
- [8] Lee Rhodes. System and method for enhanced accuracy cardinality estimation, September 24 2015. US Patent 20,150,269,178.
- [9] Salvatore Sanfilippo. Redis new data structure: The HyperLogLog. <http://antirez.com/news/75>, 2014.
- [10] Aiyu Chen, Jin Cao, Larry Shepp, and Tuan Nguyen. Distinct counting with a self-learning bitmap. *Journal of the American Statistical Association*, 106(495):879–890, 2011.
- [11] Aiyu Chen, Jin Cao, and Lawrence E Menten. Adaptive distinct counting for network-traffic monitoring and other applications, January 6 2015. US Patent 8,931,088.
- [12] Philippe Flajolet and G. Nigel Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31(2):182 – 209, 1985.
- [13] Kyu-Young Whang, Brad T. Vander-Zanden, and Howard M. Taylor. A linear-time probabilistic counting algorithm for database applications. *ACM Transactions on Database Systems*, 15(2):208–229, June 1990.
- [14] Philippe Jacquet and Wojciech Szpankowski. Analytical depoissonization and its applications. *Theoretical Computer Science*, 201(1):1–62, 1998.
- [15] Marianne Durand and Philippe Flajolet. Loglog counting of large cardinalities. In *Proceedings of the 11th Annual European Symposium on Algorithms*, pages 605–617, Budapest, Hungary, September 2003.
- [16] George Casella and Roger L Berger. *Statistical inference*. Duxbury, Pacific Grove, CA, USA, 2nd edition, 2002.
- [17] Peter Clifford and Ioana A. Cosma. A statistical analysis of probabilistic counting algorithms. *Scandinavian Journal of Statistics*, 39(1):1–14, 2012.

- [18] Daniel Ting. Towards optimal cardinality estimation of unions and intersections with sketches. In *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge and Data Mining*, pages 1195–1204, San Francisco, CA, USA, August 2016.
- [19] Reuven Cohen, Liran Katzir, and Aviv Yehezkel. A minimal variance estimator for the cardinality of big data set intersection. *arXiv preprint arXiv:1606.00996*, 2016.
- [20] William H. Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge University Press, 2007.
- [21] Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Patrick Alken, Michael Booth, Fabrice Rossi, and Rhys Ulerich. GNU scientific library reference manual. <http://www.gnu.org/software/gsl>, 2016.