Analysis of a Category of Probabilistic Cardinality Estimation Algorithms

Jinyi Wang

***Abstract*—Accurately assessing the count of unique elements within voluminous data streams remains a critical task in data analytics. The pioneering Flajolet-Martin algorithm and its descendants, such as HyperLogLog, have pioneered the arena of probabilistic counting techniques. However, there has been ongoing discussion regarding the impact of hash function value distribution on the performance of these algorithms. This study disputes the widely held belief that the accuracy of cardinality estimation algorithms is highly dependent on the distribution of hash values. We demonstrate that, for a broad spectrum of estimators, the minimum possible variance, as dictated by the Cramér-Rao lower bound, is actually unaffected by the choice of hash value distribution in extreme value-based counters. To validate our theoretical assertions, we present a novel sketching method called Pareto sketching. Our empirical tests show that this method delivers precision on par with the established exponential sketching methods. Our work not only simplifies the design of future sketching algorithms but also opens new directions for research in cardinality estimation that are not constrained by distributional choices.**

***Index Terms*—Cardinality estimation, streaming algorithms, data sketch, distinct elements problem**

I. Introduction

COUNTING the number of distinct elements in a data stream is a fundamental problem in data systems. A wide variety of applications, especially in the field of data mining and information retrival rely on efficient counting of distinct elements, such as duplicate document identification [1], graph neighbourhood function estimation [2] and genome classification [3]. However, it has been proved that any exact algorithm for this problem will require storage space linear to the number of distinct elements [4], which is infeasible for massive data streams of billions of elements. The need for efficient approximate algorithms for distinct elements problem naturally arises as massive data streams becomes more and more common in realistic applications.

In 1985, Flajolet and Martin proposed the first probabilistic algorithm known as the Flajolet-Martin algorithm [5] to approximate the number of distinct elements in the stream, with each item being examined only one pass. They pioneered utilizing deterministic hash functions to associate random variables with elements. In Flajolet-Martin algorithm, a uniform hash function h_i maps an element e to a L-bit binary string $h(e) \in \{0,1\}^L$, the largest number of consecutive zeros starting from the least significant bit $\rho(h(e))$ among all the elements is maintained. Obviously, as there are more distinct elements in the stream, a hash string with more consecutive zeros is more likely to occur. The subsequent work of Flajolet-Martin algorithm, LogLog algorithm proposed by Flajolet *et, al.* [6] reduces the space required for the desired estimation accuracy to $\log_2 \log_2 N + O(1)$, where N is the number of distinct elements in the stream. HyperLogLog [7] further improves the estimator used by LogLog, and currently still serves as the state-of-the-art cardinality estimation algorithm in many applications. Exponential sketching [8] proposed by Lemiesz utilizes continuous hash functions instead of discrete bit-patterns for cardinality estimation. Even though not as memory efficient as HyperLogLog and its variants, exponential sketching is capable of estimating cardinality of multisets in additional to binary sets.

Most existing methods for approximating cardinality make use of the minimum or maximum of random variables of elements. Let the hash value for different elements $h(e_j)$, where $j \in \{1, 2, \cdots, N\}$ is the index of element, be independent and identically distributed random variables drawn from some distribution with cumulative distribution function F. The expression of the cumulative distribution function of $\max_j h(e_j)$ can be easily derived by $F^*(x) = F(x)^N$. The distribution of the maximum is thus a statistic related to the number of distinct elements N as long as the distribution F is not degenerate. Question naturally arises on what effect the choice of distribution of hash values will have on the accuracy of estimation. According to the well-known Cramér-Rao bound on the variance of unbiased and biased estimators, the minimum variance that an estimator can archive on estimating some parameter θ is associated with the Fisher information the distribution function carries about θ. The question now becomes what effect different distributions of hash values will have on the accuracy of estimation. In this paper we have proved that the minimum attainable variance is generality unrelated with the distribution of hash values. We proposed a new method named Pareto sketching and proved that Pareto sketching achieves the same estimation accuracy as exponential sketching proposed in [8].

The remaining part of this article is organized as follows: In Section II we first formulate the problem, define what specifically extreme value counters are and then prove that the Cramér-Rao lower bound of variance of estimation are generally unrelated to distributions used in counters. In Section III we particularly analyze an existing exponential sketching scheme utilizing our theory of extreme value counters. A new sketching scheme named Pareto sketching is also proposed. In Section IV experiments are conducted to support our theoretical analysis and to prove our Pareto sketching has equivalent performance to exponential sketching.

II. The Theory of Extreme Value Counters

A. Problem Formulation

Consider a data stream $\mathcal{S} = (e_{s_1}, e_{s_2}, \cdots, e_{s_M})$ of length M consisting of N distinct elements $e_1, e_2, \cdots, e_N \in \mathcal{U}$, where $s_1, s_2, \cdots, s_M \in \{1, 2, \cdots N\}$ and $\mathcal{U}$ is the universe of elements.

Definition II.1 (hash function)**.** A hash function with distribution f is a bijection $\mathcal{U} \mapsto \mathcal{H}$, mapping elements in the universe $\mathcal{U}$ to a collection $\mathcal{H}$ of $|\mathcal{U}|$ independent and identically distributed random variables with law f.

Definition II.2 (maximum counter)**.** Without loss of generality, an maximum counter is the maximum order statistic $h^* = \max h(\mathcal{S}) = \max(h(e_{s_1}), h(e_{s_2}), \cdots, h(e_{s_M}))$.

Definition II.3 (minimum counter)**.** The concept of minimum counters can be defined in the same way as maximum counters. A minimum counter is the minimum order statistic $h\dagger = \min h(\mathcal{S}) = \max(h(e_{s_1}), h(e_{s_2}), \cdots, h(e_{s_M}))$.

The theory for maximum counters can be easily applied to minimum counters with slight modification, and our argument will thus mainly focus on maximum counters.

B. Cramér-Rao Bound for Estimators

The following important theorem by Cramér and Rao gives bound on the minimum variance of unbiased estimators:

Theorem II.1 (Cramér-Rao Bound)**.** *Suppose that some unknown parameter θ is going to be estimated from k independent observations of the random variable X with law $f(x;\theta)$, than the variance of any estimation $\hat{\theta}$ is bounded by the reciprocal of the Fisher information $I(\theta)$:*

$$\mathbf{D}_X\left[\hat{\theta}\right] \geq \frac{1}{I(\theta; f)}$$

where the Fisher information is defined by

$$I(\theta) = k\mathbf{E}_X\left[\left(\frac{\partial \log f(X;\theta)}{\partial\theta}\right)^2\right].$$

and the bound is attained by an efficient estimator if exist. Moreover, if $\log f$ is twice differentiable w.r.t. θ, and following three regularity conditions

- *f is differentiable w.r.t. θ almost everywhere.*
- *f can be differentiated under integral w.r.t. θ.*
- *The support of f does not depend on θ.*

are met, we can use the following expression for the Fisher information, which is less complex in many situations:

$$I(\theta; f) = k\mathbf{E}_X\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta^2}\right].$$

Recall that the cumulative distribution function of the maximum counter can be expressed in terms of the cumulative distribution function of the hash function, the Cramér-Rao Bound for the maximum counter can thus be derived. Let the cumulative distribution function and probability density function of the maximum counter h be denoted F^* and f^*, respectively, and for the Fisher information of θ we have

$$\begin{aligned} I(\theta) &= k\mathbf{E}_{h^*}\left[\left(\frac{\partial \log f^*(h^*;\theta)}{\partial\theta}\right)^2\right] \\ &= k\mathbf{E}_{h^*}\left[\left(\frac{\partial}{\partial\theta}\log\frac{\partial F(h^*;\theta)^n}{\partial h}\right)^2\right] \\ &= k\mathbf{E}_{h^*}\left[\left(\frac{\partial}{\partial\theta}\log\left(nf(h^*;\theta)F(h^*;\theta)^{n-1}\right)\right)^2\right]. \end{aligned}$$

Furthermore, in the task of estimating the number of distinct elements, the parameter to be estimated is just n, and function F and f are independent of n. By substituting θ with n and eliminating θ from the expression of f and F, there is

$$\begin{aligned} I(n) &= k\mathbf{E}_{h^*}\left[\left(\frac{\partial}{\partial n}\log\left(nf(h^*)F(h^*)^{n-1}\right)\right)^2\right] \\ &= k\mathbf{E}_{h^*}\left[\left(\frac{\partial}{\partial n}\left(\log n + \log f(h^*) + \log F(h^*)^{n-1}\right)\right)^2\right] \\ &= k\mathbf{E}_{h^*}\left[\left(\frac{1}{n} + \log F(h^*)\right)^2\right]. \end{aligned} \tag{1}$$

If f^* meets the regularity conditions, the expression for the Fisher information can be further simplified:

$$\begin{aligned} I(n) &= k\mathbf{E}_{h^*}\left[\frac{\partial^2 \log f^*(h^*;n)}{\partial n^2}\right] \\ &= k\mathbf{E}_{h^*}\left[\frac{\partial^2}{\partial n^2}\left(\log n + \log f(h^*) + \log F(h^*)^{n-1}\right)\right] \\ &= \frac{k}{n^2} \end{aligned}$$

Here we successfully derived the Cramér-Rao bound of cardinality estimation for maximum counters:

Theorem II.2 (Cramér-Rao bound for maximum counters)**.** *Given a maximum counter h^* where the underlying hash function h has cumulative distribution function F, the minimum variance of estimators based on h^* is bounded by*

$$\mathbf{D}_{h^*}\left[\hat{\theta}(h^*)\right] \geq \mathbf{E}_{h^*}\left[\left(\frac{1}{n} + \log F(h^*)\right)^2\right]$$

and in most situations where the regularity conditions are met, the bound is independent of the distribution of h:

$$\mathbf{D}_{h^*}\left[\hat{\theta}(h^*)\right] \geq \frac{n^2}{k}$$

III. Cardinality Estimation Based on Extreme Value Counters

Now we study particular cases of cardinality estimation algorithms based on extreme value counts using the theorems in section II.

A. Exponential Sketching

The exponential sketching scheme [8] proposed by Lemiesz choose the exponential distribution as the distribution of h. The minimum order statistic from the exponential distribution has a simple closed-form expression. Let there be exponentially distributed random variables with different rate parameters $h(e_1) \sim \text{Exp}(\lambda_1), h(e_2) \sim \text{Exp}(\lambda_2), ..., h(e_n) \sim \text{Exp}(\lambda_n)$, their smallest order statistic $h^\dagger$ are also exponentially distributed, whose rate parameter equal to the sum of the rate parameter of individual exponential variables. Thus an approximation to the cardinality can be obtained by estimating the rate parameter of X^*. As exponential variables with rate parameter λ can be costlessly generated by dividing an standard exponential variable by λ, exponential sketching can be also used for cardinality estimation of multisets.

Exponential Sketching makes use of a minimum counter, which is not the same as maximum counters discoursed in our pervious theorem II.2. Actually, the corresponding theorem for minimum counters can be derived in the same way as in II.2.

Theorem III.1 (Cramér-Rao bound for minimum counters)**.** *The Cramér-Rao bound for minimum counter $h^\dagger$ is also*

$$\mathbf{D}_{h^\dagger}\left[\hat{\theta}(h^\dagger)\right] \geq \frac{n^2}{k}$$

as long as the regularity conditions are met.

Proof. The cumulative function of the smallest order statistic $h^\dagger$ from independent and identically distributed random variables h can be written in terms of the cumulative distribution function of h:

$$F^\dagger = 1 - (1 - F)^n.$$

Thus the same bound can be derived as follows:

$$\begin{aligned} I(\theta) &= k\mathbf{E}_{h^\dagger}\left[\left(\frac{\partial \log f^\dagger(h^\dagger; n)}{\partial \theta}\right)^2\right] \\ &= k\mathbf{E}_{h^\dagger}\left[\left(\frac{\partial}{\partial \theta}\log\left(nf(h^\dagger)\left(1 - F(h^\dagger)\right)^{n-1}\right)\right)^2\right] \\ &= k\mathbf{E}_{h^\dagger}\left[\left(\frac{1}{n} + \log F(h^\dagger)\right)^2\right]. \end{aligned} \tag{2}$$

When the regularity conditions are met, the Fisher information $I(n)$ is equal to

$$\begin{aligned} &k\mathbf{E}_{h^\dagger}\left[\left(\frac{\partial}{\partial n}\log\left(nf(h^\dagger)\left(1 - F(h^\dagger)\right)^{n-1}\right)\right)^2\right] \\ =&k\mathbf{E}_{h^\dagger}\left[\frac{\partial^2}{\partial n^2}\left(\log n + \log f(h^\dagger) + \log\left(1 - F(h^\dagger)\right)^{n-1}\right)\right] \\ =&\frac{k}{n^2} \end{aligned}$$

□

Hence, the bound for minimum counters and maximum counters are the same as expected. Obviously the regularity conditions are met for the exponential distribution, or alternatively we could use the unsimplified version (2) to obtain the same result:

$$\begin{aligned} I(\theta) =&k\mathbf{E}_{h^\dagger}\left[\left(\frac{1}{n} + \log F(h^\dagger)\right)^2\right] \\ =&k\int_0^{+\infty}\left(\frac{1}{n} + \log F(h^\dagger)\right)^2 \mathrm{d}F^\dagger \\ =&k\int_0^{+\infty} ne^{-nh^\dagger}\left(\frac{1}{n} - h^\dagger\right)^2 \mathrm{d}h^\dagger \\ =&k\int_0^{+\infty} ne^{-nh^\dagger}\left(\frac{1}{n^2} - \frac{2h^\dagger}{n} + {h^\dagger}^2\right)\mathrm{d}h^\dagger \\ =&k\left(\frac{1}{n^2} - \frac{2}{n^2} + \frac{2}{n^2}\right) \\ =&\frac{k}{n^2}. \end{aligned}$$

Even though the bound on estimator variance has been proved, the variance of estimation will depend on the estimator used in practice. Unfortunately, there is no efficient estimator exists for estimating the rate parameter of exponential distribution. The Minimum-Variance Unbiased Estimator for rate parameter is

$$\hat{\lambda} = (K-1)\left(\sum_{k=1}^{K} h_k^\dagger\right)^{-1}$$

where K is the number of observations, and the variance of the Minimum-Variance Unbiased Estimator is

$$\mathbf{D}\left[\hat{\lambda}\right] = \frac{n^2}{K-2}$$

which indicates that this estimator is asymptotically efficient.

B. Pareto Sketching

We now propose a novel sketching scheme which has estimation accuracy equivalent to the exponential sketching based on our argument above. Another distribution with simple closed-form smallest order statistic is the Pareto distribution. Given n independent and identically distributed Pareto random variables, namely $h(e_1) \sim \text{Pareto}(\alpha_1), \cdots, h(e_n) \sim \text{Pareto}(\alpha_n)$ with the same scaleλ and different shapes α_i, a similar law holds as in the exponential setting. Their smallest order statistic $h^\dagger$ follows the Pareto distribution $\text{Pareto}(\lambda, \alpha_1 + \alpha_2 + \cdots + \alpha_n)$. For Pareto distribution we can estimate its shape parameter α by

$$\hat{\alpha} = (K-1)\left(\sum_{k=1}^{K} \log h_k^\dagger\right)^{-1}$$

which is the Minimum-Variance Unbiased Estimator. The Minimum-Variance Unbiased Estimator for the Pareto distribution has the same variance as that for the exponential distribution:

$$\mathbf{D}\left[\hat{\alpha}\right] = \frac{n^2}{K-2}$$

which indicates that $\hat{\alpha}$ is also an asymptotically efficient estimator as well as $\hat{\lambda}$.

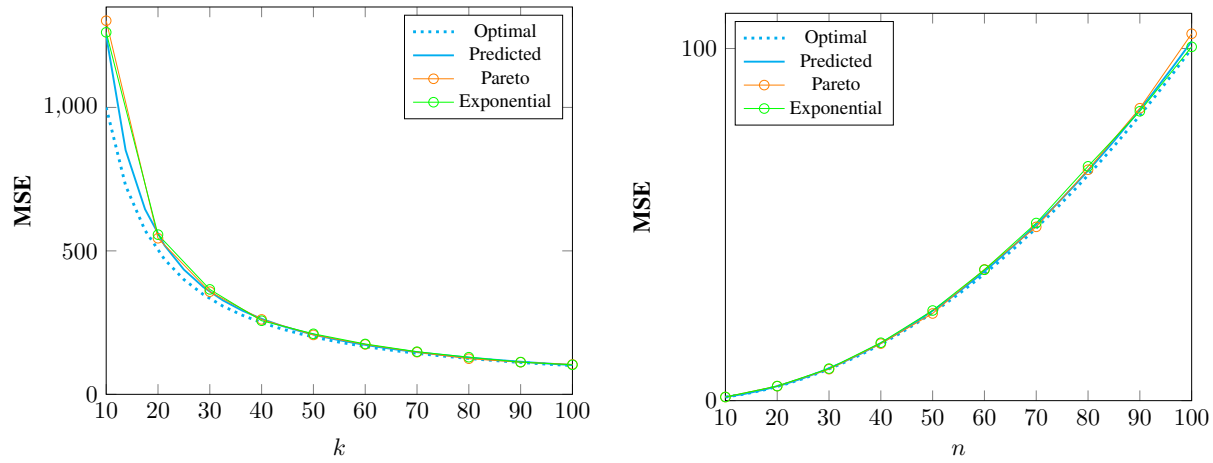

Fig. 1. MSE of estimation with varying K and N, where the optimal Cramér-Rao Bound, variance and evaluated results of Pareto and exponential sketching are shown.

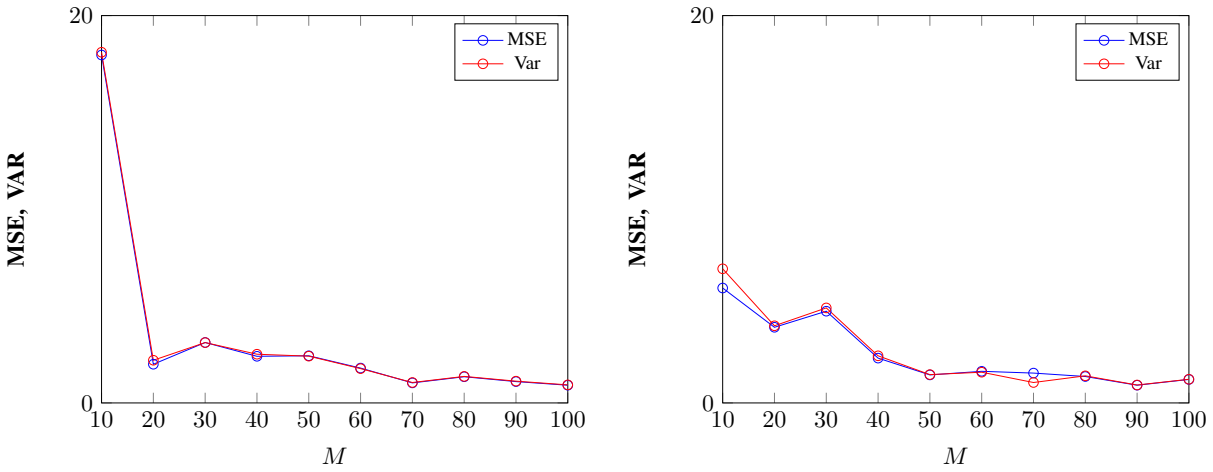

Fig. 2. The MSE and Var of Pareto and exponential sketching with K and N fixed and varying M.

IV. Experimental Evaluation

We make comparisions between the performance of both exponential sketching and Pareto sketching and predictions based on our theoretical argument to show that our theoretical bound coincides well with simulation. The metrics used in experiments are on the Mean Square Error (MSE) and the Variance (Var). Assuming that M independent experiments are conducted, the Mean Square Error is defined by

$$\text{MSE} = \frac{1}{M}\sum_{m=1}^{M}(\hat{n}_m - n)^2$$

and the Variance is defined by

$$\text{Var} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{n}_m - \bar{n})^2$$

A. Experiment Settings

In the experiments we study the influence of three different parameters, the number of observations K, the number of distinct elements N, and the number M of independent experiments whose metrics get averaged. The random variables used for estimation are generated by transforming uniform hashes from the MurMurHash3 algorithm [9].

B. The Effect of K and N on MSE

In Fig. 1 we plot the average MSE of $10,000$ independent runs, varying the parameter K and N, respectively. We evaluated the algorithms first with K varying between 10 and 100 and N fixed to 100, and then with N varying between 10 and 100 and K fixed to 100. As the number of observations K increases, the MSE goes inversely proportional to K, and the MSE goes quadraticly as the number of distinct elements N grows. We plot the predicted variance of the estimator (Predicted) in blue lines and the Cramér-Rao bound (Optimal) in blue dotted lines. We found that the predicted variance of the estimator fits simulated results well, and converges to the Cramér-Rao bound as K increases.

C. Relationship between MSE and Var

The Mean Square Error (MSE) could be decomposed into variance and the square of bias:

$$\text{MSE} = \text{Var} + \text{Bias}^2$$

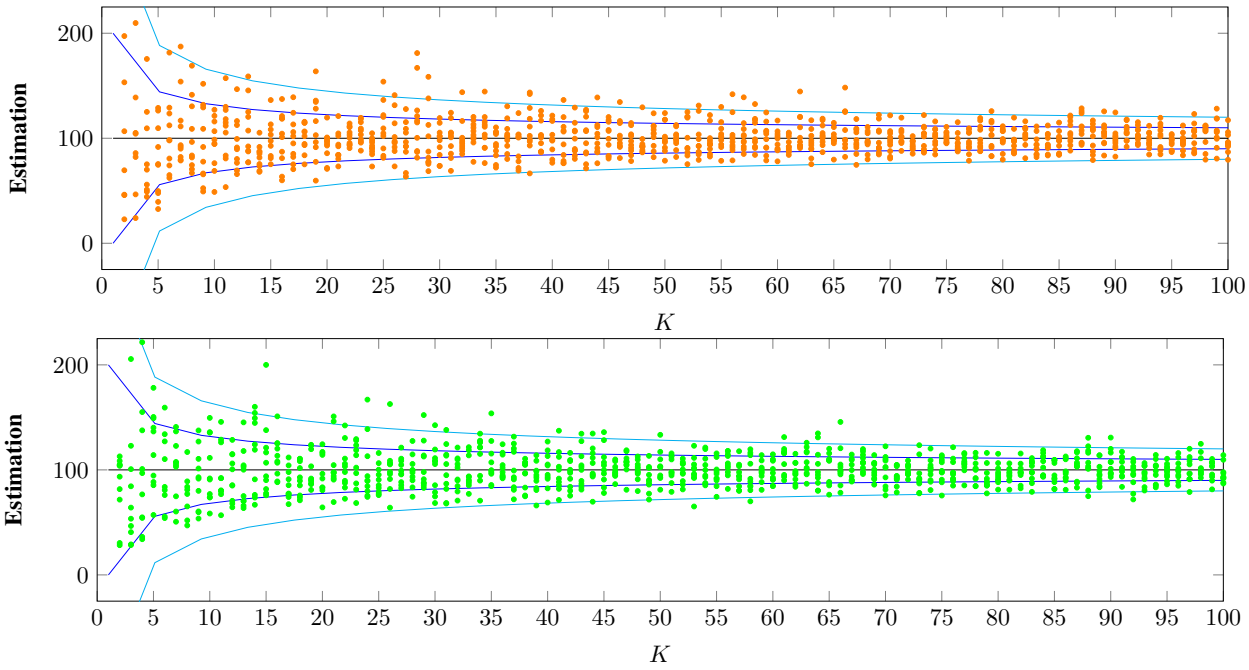

Fig. 3. Point estimations made by Pareto and exponential sketching with different K is plotted. As K grows, the estimation are more centralize towards N. The black, blue and light blue curves are the expectation, range of the standard deviation, and range of twice the standard deviations, respectively.

which is commonly known as the Bias-Variance Decomposition. In Pareto and exponential sketching which are both unbiased estimation methods, the MSE should coincide with Var. We plotted the MSE and Var of estimation with varying number of independent experiments M with fixed $K = 100$ and $N = 100$ in Fig. 2.

D. Visualization of Point Estimations

To demonstrate how the variance of estimation is reduced as the number of observations K grows in a more intuitive way, we visualize the estimations as scatters. Estimation of 10 independent runs with fixed $N = 100$ are plotted, and it can easily be observed that the estimations get centralized towards N as K grows. There are also curves for predicted standard deviation (Std, σ), which is

$$\sigma = \sqrt{\mathbf{D}} = n\sqrt{\frac{1}{K-2}}$$

and twice standard deviation. It is worth mentioning that most of the estimations lies within 2σ of the expectation N following the Central Limit Theorem.

V. Conclusion

To sum up, our study tackled the common belief that the choice of hash value distributions in cardinality estimation methods has a significant impact on the precision of the estimates. Contrary to this belief, our findings reveal that the Cramér-Rao lower bound, which dictates the best possible accuracy of any estimator, does not depend on the specific distribution of hash values employed in extreme value counters.

The implication of our research is twofold. First, it simplifies the process of designing new sketching algorithms by showing that the focus can be shifted away from the distribution of hash values. Second, it presents Pareto sketching as an example of a technique that is not only effective but also stands as evidence supporting our theoretical assertion.

Pareto sketching, the method we introduced, performs on par with the existing exponential sketching technique. This serves as a practical demonstration that the efficiency of cardinality estimation can be maintained without the need to tailor the hash value distribution to the estimator.

In essence, our work guides future endeavors in the realm of data stream analysis away from distribution-dependent designs towards potentially more critical aspects that contribute to estimator performance. By detangling the relationship between hash value distributions and the Cramér-Rao lower bound, we pave the way for more versatile and theoretically grounded approaches to cardinality estimation.

References

[1] A. Z. Broder, "Identifying and filtering near-duplicate documents," in *Combinatorial Pattern Matching*, R. Giancarlo and D. Sankoff, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–10.

[2] P. Boldi, M. Rosa, and S. Vigna, "HyperANF: Approximating the Neighbourhood Function of Very Large Graphs on a Budget," in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 625–634. [Online]. Available: https://doi.org/10.1145/1963405.1963493

[3] F. P. Breitwieser, D. N. Baker, and S. L. Salzberg, "KrakenUniq: confident and fast metagenomics classification using unique k-mer counts," *Genome Biology*, vol. 19, no. 1, p. 198, Nov 2018. [Online]. Available: https://doi.org/10.1186/s13059-018-1568-0

[4] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, ser. STOC

'96. New York, NY, USA: Association for Computing Machinery, 1996, p. 20–29. [Online]. Available: https://doi.org/10.1145/237814.237823

[5] P. Flajolet and G. N. Martin, "Probabilistic counting algorithms for data base applications," *J. Comput. Syst. Sci.*, vol. 31, no. 2, p. 182–209, sep 1985. [Online]. Available: https://doi.org/10.1016/0022-0000(85)90041-8

[6] M. Durand and P. Flajolet, "Loglog counting of large cardinalities," in *Algorithms - ESA 2003*, G. Di Battista and U. Zwick, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 605–617.

[7] P. Flajolet, É. Fusy, O. Gandouet, and F. Meunier, "HyperLogLog: the analysis of a near-optimal cardinality estimation algorithm," in *AofA: Analysis of Algorithms*, ser. DMTCS Proceedings, P. Jacquet, Ed., vol. DMTCS Proceedings vol. AH, 2007 Conference on Analysis of Algorithms (AofA 07). Juan les Pins, France: Discrete Mathematics and Theoretical Computer Science, Jun. 2007, pp. 137–156. [Online]. Available: https://inria.hal.science/hal-00406166

[8] J. Lemiesz, "On the algebra of data sketches," *Proc. VLDB Endow.*, vol. 14, no. 9, p. 1655–1667, may 2021. [Online]. Available: https://doi.org/10.14778/3461535.3461553

[9] A. Austin, Mar 2008. [Online]. Available: https://tanjent.livejournal.com/756623.html