Composable Sketches for Functions of Frequencies: Beyond the Worst Case

Edith Cohen [1 2] **Ofir Geri** [3 4] **Rasmus Pagh** [1 5 6]

Abstract

Recently there has been increased interest in using machine learning techniques to improve classical algorithms. In this paper we study when it is possible to construct compact, composable sketches for weighted sampling and statistics estimation according to functions of data frequencies. Such structures are now central components of large-scale data analytics and machine learning pipelines. However, many common functions, such as thresholds and pth frequency moments with $p > 2$, are known to require polynomial size sketches in the worst case. We explore performance beyond the worst case under two different types of assumptions. The first is having access to noisy *advice* on item frequencies. This continues the line of work of Hsu et al. (ICLR 2019), who assume predictions are provided by a machine learning model. The second is providing guaranteed performance on a restricted class of input frequency distributions that are better aligned with what is observed in practice. This extends the work on heavy hitters under Zipfian distributions in a seminal paper of Charikar et al. (ICALP 2002). Surprisingly, we show analytically and empirically that "in practice" small polylogarithmic-size sketches provide accuracy for "hard" functions.

1. Introduction

Composable sketches, also known as mergeable summaries (Agarwal et al., 2013), are data structures that support summarizing large amounts of distributed or streamed data with small computational resources (time, communication, and space). Such sketches support processing additional data elements and merging sketches of multiple datasets to obtain a sketch of the union of the datasets. This design is suitable for working with streaming data (by processing elements as they arrive) and distributed datasets, and allows us to parallelize computations over massive datasets. Sketches are now a central part of managing large-scale data, with application areas as varied as federated learning (McMahan et al., 2017) and statistics collection at network switches (Liu et al., 2016; 2019).

The datasets we consider consist of *elements* that are key-value pairs (x, v) where $v \geq 0$. The frequency w_x of a key x is defined as the sum of the values of elements with that key. When the value of all elements is 1, the frequency is simply the number of occurrences of a key in the dataset. Examples of such datasets include search queries, network traffic, user interactions, or training data from many devices. These datasets are typically distributed or streamed.

Given a dataset of this form, one is often interested in computing statistics that depend on the frequencies of keys. For example, the statistics of interest can be the number of keys with frequency greater than some constant (threshold functions), or the second frequency moment ($\sum_x w_x^2$), which can be used to estimate the skew of the data. Generally, we are interested in statistics of the form

$$\sum_x L_x f(w_x) \tag{1}$$

where f is some function applied to the frequencies of the keys and the coefficients L_x are provided (for example as a function of the features of the key x). An important special case, popularized in the seminal work of (Alon et al., 1999), is computing the f-statistics of the data: $\|f(\boldsymbol{w})\|_1 = \sum_x f(w_x)$.

One way to compute statistics of the form (1) is to compute a random sample of keys, and then use the sample to compute estimates for the statistics. In order to compute low-error estimates, the sampling has to be weighted in a way that depends on the target function f: each key x is weighted by $f(w_x)$. Since the problem of computing a weighted sample is more general than computing f-statistics, our focus in this work will be on composable sketches for weighted sampling according to different functions of frequencies.

[1]Google Research, Mountain View, CA [2]Department of Computer Science, Tel Aviv University, Israel [3]Stanford University, CA, USA [4]Most of this work was done while interning at Google Research. [5]BARC, Denmark [6]IT University of Copenhagen, Denmark. Correspondence to: Edith Cohen <edith@cohenwang.com>, Ofir Geri <ofirgeri@cs.stanford.edu>, Rasmus Pagh <pagh@itu.dk>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

The tasks of sampling or statistics computation can always be performed by first computing a table of key and frequency pairs (x, w_x). But this aggregation requires a data structure of size (and in turn, communication or space) that grows linearly with the number of keys whereas ideally we want the size to grow at most polylogarithmically. With such small sketches we can only hope for approximate results and generally we see a trade-off between sketch size, which determines the storage or communication needs of the computation, and accuracy.

When estimating statistics from samples, the accuracy depends on the sample size and on how much the sampling probabilities "suit" the statistics we are estimating. In order to minimize the error, the sampling probability of each key x should be (roughly) proportional to $f(w_x)$. This leads to a natural and extensively-studied question: for which functions f can we design efficient sampling sketches?

The literature and practice are ripe with surprising successes for sketching, including small (polylogarithmic size) sketch structures for estimating the number of distinct elements (Flajolet & Martin, 1985; Flajolet et al., 2007) ($f(w) = I_{w>0}$), frequency moments ($f(w) = w^p$) for $p \in [0, 2]$ (Alon et al., 1999; Indyk, 2001), and computing ℓ_p heavy hitters (for $p \leq 2$, where an ℓ_p ε-heavy hitter is a key x with $w_x^p \geq \varepsilon \|\boldsymbol{w}\|_p^p$) (Misra & Gries, 1982; Charikar et al., 2002; Manku & Motwani, 2002; Cormode & Muthukrishnan, 2005; Metwally et al., 2005). Here I_σ is the indicator function that is 1 if the predicate σ is true, and 0 otherwise. A variety of methods now support sampling via small sketches for rich classes of functions of frequencies (Cohen, 2018; McGregor et al., 2016; Jayaram & Woodruff, 2018; Cohen & Geri, 2019), including the moments $f(w) = w^p$ for $p \in [0, 2]$ and the family of concave sublinear functions.

The flip side is that we know of lower bounds that limit the performance of sketches using small space for some fundamental tasks (Alon et al., 1999). A full characterization of functions f for which f-statistics can be estimated using polylogarithmic-size sketches was provided in (Braverman & Ostrovsky, 2010). Examples of "hard" functions are thresholds $f(w) = I_{w>T}$ (counting the number of keys with frequency above a specified threshold value T), threshold weights $f(w) = wI_{w>T}$, and high frequency moments $f(w) = w^p$ with $p > 2$. Estimating the pth frequency moment ($\sum_x w_x^p$) for $p > 2$ is known to require space $\Omega(n^{1-2/p})$ (Alon et al., 1999; Li & Woodruff, 2013), where n is the number of keys. These particular functions are important for downstream tasks: Threshold aggregates characterize the distribution of frequencies, and high moment estimation is used in the method of moments, graph applications (Eden et al., 2019), and for estimating the cardinality of multi-way self-joins (Alon et al., 2002) (a pth moment is used for estimating a p-way join).

Beyond the worst case. Much of the discussion of sketching classified functions into "easy" and "hard". For example, there are known efficient methods for sampling according to $f(w) = w^p$ for $p \leq 2$, while for $p > 2$, even the easier task of computing the pth moment is known to require polynomial space. However, the hard data distributions used to establish lower bounds for some functions of frequency are arguably not very realistic. Real data tends to follow nice distributions and is often (at least somewhat) predictable. We study sketching where additional assumptions allow us to circumvent these lower bounds while still providing theoretical guarantees on the quality of the estimates. We consider two distinct ways of going beyond the worst case: 1) access to *advice models*, and 2) making natural assumptions on the frequency distribution of the dataset.

For the sampling sketches described in this paper, we use a notion of *overhead* to capture the discrepancy between the sampling probabilities used in the sketch and the "ideal" sampling probabilities of weighted sampling according to a target function of frequency f. An immensely powerful property of using sampling to estimate statistics of the form (1) is that the overhead translates into a multiplicative increase in sample/sketch size, without compromising the accuracy of the results (with respect to what an ideal "benchmark" weighted sample provides). This property was used in different contexts in prior work, e.g., (Frieze et al., 2004; Cohen et al., 2009), and we show that it can be harnessed for our purposes as well. For the task of estimating f-statistics, we use a tailored definition of overhead, that is smaller than the overhead for the more general statistics (1).

Advice model. The advice model for sketching was recently proposed and studied by Hsu et al. (2019). The advice takes the form of an oracle that is able to identify whether a given key is a heavy hitter. Such advice can be generated, for example, by a machine learning model trained on past data. The use of the "predictability" of data to improve performance was also demonstrated in (Kraska et al., 2018; Indyk et al., 2019). A similar heavy hitter oracle was used in (Jiang et al., 2020) to study additional problems in the streaming setting. For high frequency moments, they obtained sketch size $O(n^{1/2-1/p})$, a quadratic improvement over the worst-case lower bound.

Here we propose a sketch for *sampling by advice*. We assume an advice oracle that returns a noisy prediction of the frequency of each key. This type of advice oracle was used in the experimental section of (Hsu et al., 2019) in order to detect heavy hitters. We show that when the predicted $f(w_x)$ for keys with above-average contributions $f(w_x)$ is approximately accurate within a factor C, our sample has overhead $O(C)$. That is, the uncertainty in the advice translates to a factor $O(C)$ increase in the sketch size but does *not* impact the accuracy.

Frequencies-functions combinations. Typically, one designs sketch structures to provide guarantees for a certain function f and any set of input frequencies $\boldsymbol{w}$. The performance of a sketch structure is then analyzed for a *worst-case* frequency distribution. The analysis of the advice model also assumes worst-case distributions (with the benefits that comes from the advice). We depart from this and study sketch performance for a *combination* (F, W, h) of a family of functions F, a family W of frequency distributions, and an overhead factor h. Specifically, we seek sampling sketches that produce weighted samples with overhead at most h with respect to $f(\boldsymbol{w})$ for *every* function $f \in F$ and frequency distribution $\boldsymbol{w} \in W$. By limiting the set W of input frequency distributions we are able to provide performance guarantees for a wider set F of functions of frequency, including functions that are worst-case hard. We particularly seek combinations with frequency distributions W that are typical in practice. Another powerful property of the combination formulation is that it provides multi-objective guarantees with respect to a multiple functions of frequency F using the same sketch (Cohen, 2015; 2018; Liu et al., 2016).

The performance of sketch structures on "natural" distributions was previously considered in a seminal paper by Charikar et al. (2002). The paper introduced the *Count Sketch* structure for heavy hitter detection, where an ℓ_2 ε-heavy hitter is defined to be a key with $w_x^2 \geq \varepsilon\|\boldsymbol{w}\|_2^2$. They also show that for Zipf-distributed data with parameter $\alpha \geq 1/2$, a count sketch of size $O(k)$ can be used to find the k heaviest keys (a worst-case hard problem) and that an ℓ_1 sample can only identify the heaviest keys for Zipf parameter $\alpha \geq 1$.

We significantly extend these insights to a wider family of frequency distributions and to a surprisingly broad class of functions of frequencies. In particular we show that all high moments ($p > 2$) are "easy" as long as the frequency distribution has an ℓ_1 or ℓ_2 ε-heavy hitter. In this case, an ℓ_1 or ℓ_2 sample with overhead $1/\varepsilon$ can be used to estimate all high moments. We also show that in a sense this characterization is tight in that if we allow all frequencies, we meet the known lower bounds. It is very common for data sets in practice to have a most frequent key that is an ℓ_1 or ℓ_2 ε-heavy hitter. This holds in particular for Zipf or approximate Zipf distributions.

Moreover, we show that Zipf frequency distributions have small *universal* sketches that apply to *any* monotone function of frequency (including thresholds and high moments). Zipf frequencies were previously considered in the advice model (Aamand et al., 2019). Interestingly, we show that for these distribution a single small sketch is effective with all monotone functions of frequency, even without advice. In these cases, universal sampling is achieved with off-the-shelf polylogarithmic-size sketches such as ℓ_p samples for $p \leq 2$ and multi-objective concave-sublinear samples (Cohen et al., 2012; Cohen, 2018; McGregor et al., 2016; Jayaram & Woodruff, 2018).

Empirical study. We complement our analysis with an empirical study on multiple real-world datasets including datasets studied in prior work on advice models (Pass et al., 2006; CAIDA, 2016; Paranjape et al., 2017; Maciá-Fernández et al., 2018). (Additional discussion of the datasets appears in the supplementary material.) We apply sampling by advice, with advice based on models from prior work or direct use of frequencies from past data. We then estimate high moments from the samples. We observe that sampling-by-advice was effective on these datasets, yielding low error with small sample size. We also observed however that ℓ_2 and ℓ_1 samplers were surprisingly accurate on these tasks, with ℓ_2 samplers generally outperforming sampling by advice. This surprisingly good performance of these simple sampling schemes is suggested from our analysis.

We compute the overhead factors for some off-the-shelf sampling sketches on multiple real-world datasets with the objectives of ℓ_p sampling ($p > 2$) and universal sampling. We find these factors to be surprisingly small. For example, the measured overhead of using ℓ_2 sampling for the objective of ℓ_p sampling is in the range $[1.2, 18]$. For universal sampling, the observed overhead is lower with ℓ_1 and with multi-objective concave sublinear samples than with ℓ_2 sampling and is in $[93, 800]$, comparing very favorably with the alternative of computing a full table. Finally, we use sketches to estimate the distribution of rank versus frequency, which is an important tool for optimizing performance across application domains (for network flows, files, jobs, or search queries). We find that ℓ_1 samples provide quality estimates, which is explained by our analytical results.

2. Preliminaries

We consider datasets where each data element is a (key, value) pair. The keys belong to a universe denoted by $\mathcal{X}$ (e.g., the set of possible users or words), and each key may appear in more than one element. The values are positive, and for each key $x \in \mathcal{X}$, we define its *frequency* $w_x \geq 0$ to be the sum of values of all elements with key x. The data elements may appear as a stream or be stored in a distributed manner. We denote the number of active keys (keys with frequency greater than 0) by n.

We are interested in sketches that produce a weighted sample of keys according to some function f of their frequencies, where roughly key x is sampled with probability proportional to $f(w_x)$. We use $f(\boldsymbol{w})$ as a shorthand for the vector of all values $f(w_x)$ (in any fixed order).

Estimates from a sample. We work with sampling schemes that produce a random subset $S \subseteq \mathcal{X}$ of the keys in the dataset. For each key $x \in S$ we have its frequency w_x and can compute its probability p_x to be sampled. From such a sample, we can compute for each key x the *inverse probability estimate* (Horvitz & Thompson, 1952) of $f(w_x)$ defined as

$$\widehat{f(w_x)} = \begin{cases} \frac{f(w_x)}{p_x} & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases} .$$

These *unbiased* per-key estimates can be summed to obtain unbiased estimates of the f-statistics of a domain $H \subset \mathcal{X}$:

$$\widehat{\sum_{x\in H} f(w_x)} := \sum_{x\in H} \widehat{f(w_x)} = \sum_{x\in S} \widehat{f(w_x)} .$$

The last equality follows because $\widehat{f(w_x)} = 0$ for keys not in the sample. We can similarly estimate other statistics that are linear in $f(w_x)$, e.g., $\sum_{x\in\mathcal{X}} L_x f(w_x)$ (for coefficients L_x).

Benchmark variance bounds. We measure performance with respect to that of a "benchmark" weighted sampling scheme where each key x is sampled with probability proportional to $f(w_x)$. Recall that for "hard" functions f these schemes can not be implemented with small sketches. These benchmark schemes include (i) *probability proportional to size* (pps) with replacement where we have k independent draws where key x is selected with probability $f(w_x)/\|f(\boldsymbol{w})\|_1$, (ii) pps without replacement (ppswor (Rosén, 1972; 1997; Cohen & Kaplan, 2008)), or (iii) priority sampling (Ohlsson, 1998; Duffield et al., 2007). With these schemes, the variance for key x is upper bounded by[1]

$$\mathsf{Var}[\widehat{f(w_x)}] \leq \frac{1}{k} f(w_x) \|f(\boldsymbol{w})\|_1 \tag{2}$$

Consequently, the variance of the sum estimator for the f-statistics of a domain H is bounded by (due to nonpositive covariance shown in earlier works):

$$\mathsf{Var}\left[\widehat{\sum_{x\in H} f(w_x)}\right] \leq \frac{1}{k} \sum_{x\in H} f(w_x) \|f(\boldsymbol{w})\|_1 . \tag{3}$$

[1]With ppswor and priority sampling, $k-2$ instead of k. We use this upper bound as the benchmark, since when the sampling probability $p_x = f(w_x)/\|f(\boldsymbol{w})\|_1$ approaches 1 (one key dominates the data), the variance $(f(w_x))^2\left(\frac{1}{p_x} - 1\right)$ approaches 0, and we cannot approximate it multiplicatively with a sampling probability that multiplicatively approximates p_x. However, when there is not just one key that dominates the data, for example, $p_x \leq 1/2$ for all x, or more generally, $p_x \leq 1 - \frac{1}{c}$, we get that $\mathsf{Var}[\widehat{f(w_x)}] \geq (f(w_x))^2 \frac{1}{cp_x}$, so the bound on the variance is almost tight.

The variance on the estimate of $\|f(\boldsymbol{w})\|_1$ is bounded by

$$\mathsf{Var}[\|\widehat{f(\boldsymbol{w})}\|_1] \leq \frac{1}{k}\|f(\boldsymbol{w})\|_1^2 .$$

With these "benchmark" schemes, if we wish to get multiplicative error bound (normalized root mean squared error) of ε for estimating $\|f(\boldsymbol{w})\|_1$ we need sample size $k = O(\varepsilon^{-2})$. We note that the estimates are also concentrated in the Chernoff sense (Duffield et al., 2007; Cohen, 2015).

We refer to the probability vector

$$p_x := \frac{f(w_x)}{\|f(\boldsymbol{w})\|_1}$$

as *pps sampling probabilities* for $f(\boldsymbol{w})$. When $f(w) = w^p$ (for $p > 0$) we refer to sampling with the respective pps probabilities as ℓ_p sampling.

Emulating a weighted sample. Let $\boldsymbol{p}$ be the base pps probabilities for $f(\boldsymbol{w})$. When we use a weighted sampling scheme with weights $\boldsymbol{q} \neq \boldsymbol{p}$ then the variance bound (3) does not apply. We will say that weighted sampling according to $\boldsymbol{q}$ *emulates* weighted sampling according to $\boldsymbol{p}$ with *overhead* h if for all k and for all H, a sample of size kh provides the variance bound (3) (and the respective concentration bounds).

Lemma 2.1. *The overhead of emulating weighted sample according to* $\boldsymbol{p}$ *using weighted sampling according to* $\boldsymbol{q}$ *is at most*

$$h(p,q) := \max_x \frac{p_x}{q_x} .$$

Proof. We first bound the variance of $\widehat{f(w_x)}$ for a key x with weighted sampling by $\boldsymbol{q}$. Consider a weighted sample of size k according to base probabilities q_x. Then for all x,

$$\begin{aligned} \mathsf{Var}[\widehat{f(w_x)}] &\leq \tfrac{1}{k}(f(w_x))^2(\tfrac{1}{q_x} - 1) \leq \tfrac{1}{k}(f(w_x))^2 \tfrac{1}{p_x}\tfrac{p_x}{q_x} \\ &\leq \tfrac{1}{k} f(w_x)\|f(\boldsymbol{w})\|_1 \tfrac{p_x}{q_x} . \end{aligned}$$

The upper bound on the variance for a domain H is:

$$\frac{1}{k}\left(\sum_{x\in H} f(w_x)\right)\|f(\boldsymbol{w})\|_1 \max_{x\in H}\frac{p_x}{q_x} . \tag{4}$$

Thus for any H, the variance bound (4) is larger by the benchmark bound (3) by at most a factor of $h(p,q)$. $\square$

Note that the inaccuracy in the probabilities (using $\boldsymbol{q}$ instead of $\boldsymbol{p}$) is compensated for by a larger sample size without compromising accuracy of the estimate.

Remark 2.2. The emulation overhead can be interpreted as providing an upper bound over all possible estimation tasks that the emulated sample could be used for. This definition of overhead is tight if we wish to transfer guarantees for all subsets H: Consider a (subset that is a single) key x and sample size k such that $q_x \leq p_x \ll 1/k$. The variance when sampling according to $\boldsymbol{q}$ is $\approx (f(w_x))^2(1/(kq_x)-1) \approx (f(w_x))^2/(kq_x) = \frac{1}{k}f(w_x)\frac{f(w_x)}{p_x}\frac{p_x}{q_x} = \frac{1}{k}f(w_x)\|f(\boldsymbol{w})\|_1\frac{p_x}{q_x}$. This is a factor of p_x/q_x larger than the variance when sampling according to $p_x = f(w_x)/\|f(\boldsymbol{w})\|_1$, which is $\approx (f(w_x))^2/(kp_x) = \frac{1}{k}f(w_x)\|f(\boldsymbol{w})\|_1$.

Overhead for estimating f-statistics. If we are only interested in estimates of the full f-statistics $\|f(\boldsymbol{w})\|_1$, the overhead reduces to the expected ratio $\mathsf{E}_{x\sim\boldsymbol{p}}[p_x/q_x]$ instead of the maximum ratio.

Corollary 2.3. *Let $\boldsymbol{p}$ be the base pps probabilities for $f(\boldsymbol{w})$. Consider weighted sampling of size k according to $\boldsymbol{q}$. Then,*

$$\mathsf{Var}[\|\widehat{f(\boldsymbol{w})}\|_1] \leq \frac{\|f(\boldsymbol{w})\|_1^2}{k}\sum_x p_x \cdot \frac{p_x}{q_x} = \frac{\|f(\boldsymbol{w})\|_1^2}{k}\mathsf{E}_{x\sim\boldsymbol{p}}[\frac{p_x}{q_x}].$$

3. The Advice Model

In this section, we assume that in addition to the input, we are provided with an oracle access to an "advice" model. When presented with a key x, the advice model returns a prediction a_x for the total frequency of x in the data. For simplicity, we assume that predictions are the same for all queries with the same key. This model is similar to the model used in a recent paper about frequency estimation (Hsu et al., 2019).

A detailed description of our sketch structure and the corresponding estimators (including proofs) is provided in the supplementary material. At a high level, our sampling sketch takes size parameters (k_h, k_p, k_u), maintains a set of at most $k_h + k_p + k_u$ keys, and collects the exact frequencies w_x for these stored keys. The primary component of the sketch is a weighted-sample-by-advice of size k_p. Our sketch stores keys according to two additional criteria in order to provide robustness to the prediction quality of the advice:

- The top-k_h keys by advice. This provides tolerance to inaccuracies in the advice for these heaviest keys. Since these keys are included with probability 1, they will not contribute to the error.
- A uniform sample of k_u keys. This allows keys that are "below average" in their contribution to $\|f(\boldsymbol{w})\|_1$ to be represented appropriately in the sample, regardless of the accuracy of the advice. This provides robustness to the accuracy of the advice on these very infrequent keys and ensures they are not undersampled. Moreover, this ensures that all active keys ($w_x > 0$), including those with potentially no advice ($a_x = 0$), have a positive probability of being sampled. This is necessary for unbiased estimation.

We provide an unbiased estimator that smoothly combines the different sketch components and provides the following guarantees:

Lemma 3.1. *Suppose the advice model is such that for some $c_p, c_u \geq 0$ and $h \geq 0$, all keys x that are active ($w_x > 0$) and not in the h largest advice values of active keys ($a_x < \{a_y \mid w_y > 0\}_{(n-h+1)}$) satisfy*

$$\frac{f(w_x)}{\|f(\boldsymbol{w})\|_1} \leq \max\{c_p\frac{f(a_x)}{\|f(\boldsymbol{a})\|_1}, c_u\frac{1}{n}\}\ .$$

Then for all $k \geq 1$, a sample with $(k_h, k_p, k_u) = (h, \lceil kc_p\rceil + 2, \lceil kc_u\rceil + 2)$ satisfies the variance bound (3) for all H.

In particular, if our advice is approximately accurate, say $f(w_x) \leq f(a_x) \leq c_p \cdot f(w_x)$, the overhead when sampling by advice is c_p.

Corollary 3.2. *Let f be such that $f(w_x) \leq f(a_x) \leq c_p f(w_x)$ then for all $k \geq 1$, with sample size $(k_h, k_p, k_u) = (0, \lceil kc_p\rceil + 2, 0)$ we have $\mathsf{Var}\left[\|\widehat{f(\boldsymbol{w})}\|_1\right] \leq \frac{1}{k}\|f(\boldsymbol{w})\|_1^2$.*

Experiments. We evaluate the effectiveness of "sampling by advice" for estimating the frequency moments with $p = 3, 7, 10$ on datasets from (Pass et al., 2006; Paranjape et al., 2017). We use advice models from prior work (Hsu et al., 2019) based on a machine learning algorithms applied to past data and advice based directly on frequencies in past data. Some representative results are reported in Figure 1 and additional results and more details are provided in the supplementary material. The results reported for sampling by advice are with $k_h = 0$ and two choices of balance between the ppswor sample based on the advice and the uniform sample: $k_p = k_u$ and $k_u = 32$. We also report performance of ppswor (ℓ_1 sampling without replacement), ℓ_2 sampling (with and without replacement), and the benchmark upper bound ($1/\sqrt{k}$).

We observe that for the third moment, sampling by advice did not perform significantly better than ppswor (and sometimes performed even worse). For higher moments, however, sampling by advice performed better than ppswor when the sample size is small. With replacement ℓ_2 sampling was more accurate than advice (without replacement ℓ_2 sampling performs best). Our analysis in the next section explains the perhaps surprisingly good performance of ℓ_1 and ℓ_2 sampling schemes.

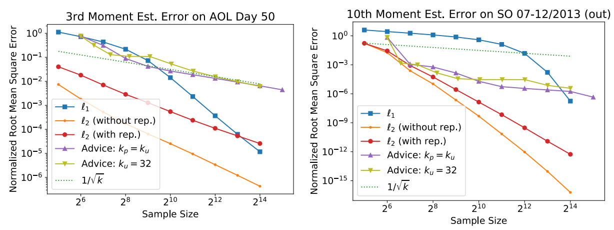

Figure 1. Estimating the third moment on the AOL dataset (with learned advice) and the tenth moment on the Stack Overflow dataset (with past frequencies).

4. Frequencies-Functions Combinations

In this section, we analyze performance for inputs that comes from restricted families of frequency distributions W. Restricting the family W of possible inputs allows us to extend the family F of functions of frequency that can be efficiently emulated with a small overhead.

Specifically, we will consider sampling sketches and corresponding *combinations* (W, F, h) of frequency vectors W, functions of frequency F, and overhead $h \geq 1$ so that for every frequency distribution $w \in W$ and frequency function $f \in F$, our sampling sketch emulates a weighted sample of $f(\boldsymbol{w})$ with overhead h. We will say that our sketch *supports* the combination (W, F, h).

Recall that emulation with overhead h means that a sampling sketch of size $h\varepsilon^{-2}$ (i.e., holding this number of keys or hashes of keys) provides estimates with NRMSE ε for $f(\boldsymbol{w})$ for *any* $f \in F$ and that these estimates are concentrated in the Chernoff bound sense. Moreover, for any $f \in F$ we can estimate statistics of the form (1) with the same guarantees on accuracy as provided by a dedicated weighted sample according to f.

We study combinations supported by off-the-shelf sampling schemes that can be implemented with small (e.g. polylogarithmic) size sketches: ℓ_q sampling for $q \leq 2$, ppswor (ℓ_1 sampling without replacement) (Gibbons & Matias, 1998; Estan & Varghese, 2002; Cohen et al., 2012), ℓ_2 samplers (with replacement) (Indyk, 2001; McGregor et al., 2016; Jayaram & Woodruff, 2018), and multi-objective concave-sublinear sampling (Cohen, 2018). We will use the notation $\boldsymbol{w} = \{w_i\}$ for dataset frequencies, where w_i is the frequency of the ith most frequent key.

We reports results of experiments on datasets listed in Table 1 with details provided in the supplementary material.

4.1. Emulating an ℓ_p Sample by an ℓ_q Sample

We express the overhead of emulating ℓ_p sampling by ℓ_q sampling ($p \geq q$) in terms of the properties of the frequency distribution. Recall that ℓ_p sampling (and estimates of p-th frequency moments) can be implemented with polylogarithmic size sketches for $p \leq 2$ but requires polynomial size sketches in the worst case when $p > 2$.

Lemma 4.1. *Consider a dataset with frequencies $\boldsymbol{w}$ (in nonincreasing order). For $p \geq q$, the overhead of emulating ℓ_p sampling by ℓ_q sampling is bounded by*

$$\frac{\left\|\frac{\boldsymbol{w}}{w_1}\right\|_q^q}{\left\|\frac{\boldsymbol{w}}{w_1}\right\|_p^p}. \tag{5}$$

Proof. The sampling probabilities for key i under ℓ_p sampling and ℓ_q sampling are $\frac{w_i^p}{\|w\|_p^p}$ and $\frac{w_i^q}{\|w\|_q^q}$, respectively. Then, the overhead of emulating ℓ_p sampling by ℓ_q sampling is

$$\max_i \frac{w_i^p/\|w\|_p^p}{w_i^q/\|w\|_q^q} = \max_i w_i^{p-q} \cdot \frac{\|w\|_q^q}{\|w\|_p^p} = w_1^{p-q} \cdot \frac{\|w\|_q^q}{\|w\|_p^p} = \frac{\left\|\frac{\boldsymbol{w}}{w_1}\right\|_q^q}{\left\|\frac{\boldsymbol{w}}{w_1}\right\|_p^p}.$$

□

We can obtain a (weaker) upper bound on the overhead, expressed only in terms of q, that applies to all $p \geq q$:

Corollary 4.2. *The overhead of emulating ℓ_p sampling using ℓ_q sampling (for any $p \geq q$) is at most $\left\|\frac{\boldsymbol{w}}{w_1}\right\|_q^q$.*

Remark 4.3. Emulation works when $p \geq q$. When $q > p$, the maximum in the overhead bound (see proof of Lemma 4.1) is incurred on the least frequent key, with frequency w_n. We therefore get a bound of $\left\|\frac{\boldsymbol{w}}{w_n}\right\|_q^q / \left\|\frac{\boldsymbol{w}}{w_n}\right\|_p^p$ and Corollary 4.2 does not apply.

4.2. Frequency Distributions with a Heavy Hitter

We show that for distributions with an ℓ_q heavy hitter, ℓ_q sampling emulates ℓ_p sampling for all $p \geq q$ with a small overhead.

Definition 4.4. *Consider frequencies $\boldsymbol{w}$. An ℓ_q ϕ-heavy hitter is defined to be a key such that $w_i^q \geq \phi \cdot \|\boldsymbol{w}\|_q^q$.*

We rephrase Corollary 4.2 in terms of a presence of a heavy hitter:

Corollary 4.5. *Let $\boldsymbol{w}$ be a frequency vector with a ϕ-heavy hitter under ℓ_q. Then for $p \geq q$, the overhead of using ℓ_q sample to emulate an ℓ_p sample is at most $1/\phi$.*

We are now ready to specify combinations (W, F, h) of frequency vectors W, functions of frequency F, and overhead $h \geq 1$ that are supported by ℓ_q sampling.

Theorem 4.6. *For any $q > 0$ and $\phi \in (0, 1]$, an ℓ_q-sample supports the combination*

$$\begin{aligned} W &:= \{\boldsymbol{w} \text{ with an } \ell_q\ \phi\text{-heavy hitter}\} \\ F &:= \overline{\{f(w) = w^p \mid p \geq q\}}_+ \\ h &:= 1/\phi\ , \end{aligned}$$

where the notation $\overline{F}_+$ is the closure of a set F of functions under nonnegative linear combinations.

In particular, if the input distribution has an ℓ_q ϕ-heavy hitter then ℓ_q sampling of size ε^{-2}/ϕ emulates an ℓ_p sampling of size ε^{-2} for any $p > q$.

Table 1. Datasets

Dataset	$n/10^6$	$\ln n$	ℓ_1 HH	ℓ_2 HH	α
SO in	2.23	14.6	0.0016, 0.0010	0.053, 0.022	1.48
SO out	2.30	14.6	0.0015, 0.0009	0.107, 0.036	1.38
AOL	10.15	16.1	0.0275, 0.0091	0.827, 0.091	0.77
CAIDA	1.07	13.9	0.0033, 0.0032	0.048, 0.046	1.35
UGR	79.38	18.2	0.1118, 0.0401	0.850, 0.109	1.35

Table 1 reports properties and the relative ℓ_1 and ℓ_2 weights of the most frequent key for our datasets. We can see that the most frequent key is a heavy hitter with $1/\phi \leq 21$ for ℓ_2 and $1/\phi \leq 625$ for ℓ_1 which gives us upper bounds on the overhead of emulating any ℓ_p sample ($p \geq 2$). Table 3 reports (for $p = 3, 7, 10$) the overhead of emulating the respective ℓ_p sample and the (smaller) overhead of estimating the pth moment. We can see that high moments can be estimated well from ℓ_2 and with a larger overhead from ℓ_1 samples.

Certified emulation. The quality guarantees of a combination (W, F, h) are provided when $\boldsymbol{w} \in W$. In practice, however, we may compute samples of arbitrary dataset frequencies $\boldsymbol{w}$. Conveniently, we are able to test the validity of emulation by considering the most frequent key in the sample: For an ℓ_q sample of size k we can compute $r \leftarrow \max_{x \in S} w_x^q / \|\boldsymbol{w}\|_q^q$ and certify that our sample emulates ℓ_p samples ($p > q$) of size kr. If kr is small, then we do not provide meaningful accuracy but otherwise we can certify the emulation with sample size kr. When the input $\boldsymbol{w}$ has an ℓ_q ϕ-heavy hitter then an ℓ_q sample of size k will include it with probability at least $1 - e^{-k\phi}$ and the result will be certified. Note that the result can only be certified if there is a heavy hitter.

Tradeoff between W and F. If $\boldsymbol{w}$ has an ℓ_q ϕ-heavy hitter then it has an ℓ_p ϕ-heavy hitter for every $p \geq q$. This means that for moments $p \geq 2$, an ℓ_2 sample supports a larger set W of frequencies than an ℓ_1 sample, including also those with an ℓ_2 ϕ-heavy hitter but not an ℓ_1 ϕ-heavy hitter. The ℓ_1 sample however supports a larger family F that includes moments with $p \in [1, 2)$. Note that for a fixed overhead and ℓ_q sampling, the set F of supported functions decreases with q whereas W increases with q.

Table 2. Supported combinations for $\mathsf{subZipf}[\alpha, c, n]$ frequencies

method	W	F	Overhead
ℓ_1 sampling	$\{\alpha \geq 2\}$	$\overline{\{f(w) = w^p \mid p \geq 1\}}_+$	$1.65c$
ℓ_1 sampling	$\{\alpha \geq 1\}$	$\overline{\{f(w) = w^p \mid p \geq 1\}}_+$	$(1 + \ln n)c$
ℓ_2 sampling	$\{\alpha \geq 1\}$	$\overline{\{f(w) = w^p \mid p \geq 2\}}_+$	$1.65c^2$
ℓ_2 sampling	$\{\alpha \geq 1/2\}$	$\overline{\{f(w) = w^p \mid p \geq 2\}}_+$	$(1 + \ln n)c^2$

4.3. Zipfian and sub-Zipfian Frequencies

Zipf distributions are a very common model for frequency distributions in practice. We explore supported combinations with frequencies that are (approximately) Zipf.

Definition 4.7. *We say that the frequencies $\boldsymbol{w}$ where $\|\boldsymbol{w}\|_0 = n$ are $\mathsf{Zipf}[\alpha, n]$ (Zipf with parameter α) if for all i, $w_i/w_1 = i^{-\alpha}$.*

Values $\alpha \in [1, 2]$ are common in practice. The best-fit Zipf parameter for the datasets we studied is reported in Table 1 and the frequency distribution (sorted by rank) is shown in Figure 2. We can see that our datasets are approximately Zipf (which would be an approximate straight line) and for all but one we have $\alpha \in [1.3, 1.5]$.

We now define a broader class of approximately Zipfian distributions.

Definition 4.8. *Frequencies $\boldsymbol{w}$ are $\mathsf{subZipf}[\alpha, c, n]$ if for all i, $\frac{w_i}{w_1} \leq ci^{-\alpha}$.*

Note that $\mathsf{Zipf}[\alpha, n]$ is sub-Zipfian with the same α and $c = 1$. We show that sub-Zipf frequencies (and in particular Zipf frequencies) have heavy hitters:

Lemma 4.9. *For $\mathsf{subZipf}[\alpha, c, n]$ frequencies, and q such that $q\alpha \geq 1$, the frequency vector has an ℓ_q $c^q \frac{1}{H_{n,\alpha q}}$-heavy hitter, where $H_{n,\alpha} := \sum_{i=1}^{n} i^{-\alpha}$ are the generalized harmonic numbers.*

Table 2 lists supported combinations that include these approximately Zipfian distributions. We see that for these approximately Zipf distributions, ℓ_1 or ℓ_2 samples emulate ℓ_p samples with small overhead.

4.4. Experiments on Estimate Quality

The overhead factors reported in Table 3 are in a sense worst-case upper bounds (for the dataset frequencies). Figure 4 reports the actual estimation error (normalized root mean square error) for high moments for representative datasets as a function of sample size. The estimates are with ppswor (ℓ_1 sample with replacement) and ℓ_2 samples with and without replacement. Additional results are reported in the supplementary material. We observe that actual accuracy is significantly better than even the benchmark bounds.

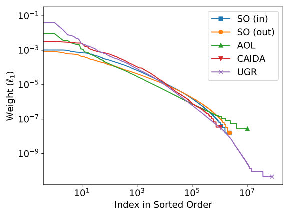

Figure 2. Freq. by rank

Table 3. Overhead

Dataset	ℓ_1 max overhead (expected overhead)			ℓ_2 max overhead (expected overhead)			concave
	3rd	10th	universal	3rd	10th	universal	universal
SO in	124.30 (42.76)	600.44 (577.57)	624.58	3.74 (1.90)	18.04 (17.36)	1.25×10^5	1672.60
SO out	299.80 (155.32)	677.56 (673.45)	681.72	4.12 (2.58)	9.30 (9.25)	4.20×10^4	1628.36
AOL	34.81 (33.45)	36.38 (36.37)	92.92	1.16 (1.12)	1.21 (1.21)	2.94×10^5	170.84
CAIDA	31.23 (18.73)	90.66 (56.20)	301.28	2.16 (1.57)	6.28 (4.03)	2.65×10^5	846.15
UGR	8.52 (8.17)	8.95 (8.95)	772.83	1.12 (1.09)	1.18 (1.18)	1.89×10^{11}	143.94

Finally we consider estimating the full distribution of frequencies, that is, the curve that relates frequency of keys to their rank. We do this by estimating the actual rank of each key in the sample (using an appropriate threshold function of frequency). Representative results are reported in Figure 3 for ppswor and for with-replacement ℓ_2 sampling (additional results are reported in the supplementary material). We used a sample of size $k = 32$ or $k = 1024$ for each set of estimates. We observe that generally the estimates are fairly accurate even with a small sample size (despite threshold function requiring large sketches in the worst case). We see that ℓ_2 samples are accurate for the frequent keys but often have no representatives from the tail whereas the without replacement ℓ_1 samples are more accurate on the tail.

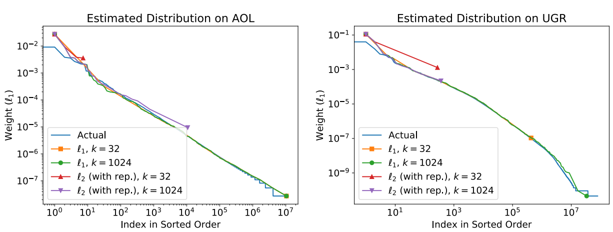

Figure 3. Actual and estimated distribution of frequency by rank. Estimates for ppswor and with-replacement ℓ_2 sampling and sample sizes $k = 32, 1024$.

4.5. Worst-Case Bound on Overhead

The overhead (5) of $\|\boldsymbol{w}/w_1\|_2^2/\|\boldsymbol{w}/w_1\|_p^p$ is the space factor increase needed for an ℓ_2 sample to emulate an ℓ_p sample on the frequencies $\boldsymbol{w}$ (and estimate accurately pth moments). A natural question is whether there is a better way to emulate an ℓ_p sampling with a polylogarithmic size sampling sketch. The following shows that in a sense an ℓ_2 sample is the best we can do:

Lemma 4.10. *An ℓ_2 sample with overhead $n^{1-2/p}$ can emulate an ℓ_p sample for any frequency vector $\boldsymbol{w}$ with support $\|\boldsymbol{w}\|_0 = n$.*

Proof. $\max_{\boldsymbol{w}} \frac{\|\boldsymbol{w}/w_1\|_2^2}{\|\boldsymbol{w}/w_1\|_p^p} \leq n^{1-2/p}$. □

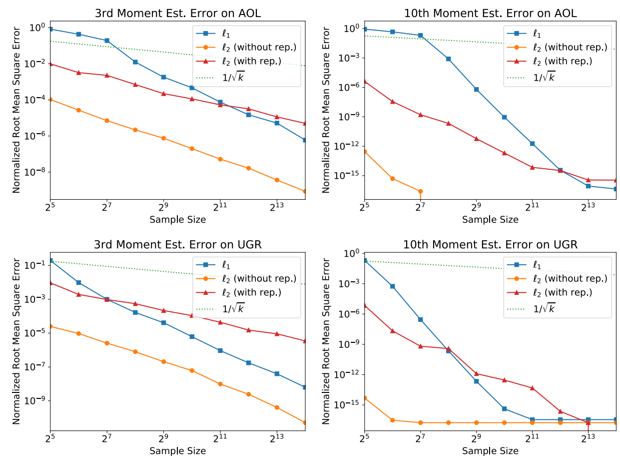

Figure 4. Estimating 3rd and 10th moments on various datasets from ppswor (ℓ_1 without replacement) and ℓ_2 samples (with and without replacement). The error is averaged over 50 repetition.

This matches upper bounds on sketch size attained with dedicated sketches for pth-moment estimation (Indyk & Woodruff, 2005; Andoni et al., 2011) and the worst-case lower bound of $\tilde{\Omega}(n^{1-2/p})$ (Alon et al., 1999; Li & Woodruff, 2013).

5. Universal Samples

We study combinations where the emulation is *universal*, that is, F includes the set M of all monotone non-decreasing functions of frequency. Interestingly, there are sampling probabilities that provide universal emulation for any $\boldsymbol{w}$:

Lemma 5.1. *(Cohen, 2015) Consider the probabilities $\boldsymbol{q}$ where the ith most frequent key has $q_i = \frac{1}{iH_n}$. Then a weighted sample by $\boldsymbol{q}$ is a universal emulator with overhead at most H_n.*

This universal sampling, however, can not be implemented with small (polylogarithmic size) sketches. This because M includes functions that require large (polynomial size) sketches such as thresholds ($I_{w \geq T}$ for some T) and high moments ($p > 2$). We therefore aim for small sketches that provide universal emulation to a restricted W.

For particular sampling probabilities $\boldsymbol{q}$ and frequencies $\boldsymbol{w}$ we consider the *universal emulation overhead* to be the overhead factor that will allow the sample to emulate weighted sampling with respect to $f(\boldsymbol{w})$ for any $f \in M$.

$$\max_{f\in M} \max_i f(w_i)/(\|f(\boldsymbol{w})\|_1 q_i) \tag{6}$$

Interestingly, the universal emulation overhead of $\boldsymbol{q}$ does not depend on the particular $\boldsymbol{w}$.

Lemma 5.2. *The universal emulation overhead of $\boldsymbol{q}$ is*

$$\max_i 1/(iq_i)$$

and is always at least H_n. This is tight even when W contains a single $\boldsymbol{w}$, as long as frequencies are distinct ($w_i > w_{i+1}$ for all i).

We can similarly consider for sampling probabilities $\boldsymbol{q}$ the *universal estimation overhead* which is the overhead needed for estimating all (full) monotone f-statistics. As discussed in Section 2, the estimation is a weaker requirement than emulation (only applies to the full statistics) and hence for any particular $\boldsymbol{q}$ the estimation overhead can be lower than the emulation overhead. The estimation overhead, however, is still at least H_n.

Lemma 5.3. *The universal estimation overhead for estimating all monotone f-statistics for $\boldsymbol{q}$ is*

$$\max_i \frac{1}{i^2}\sum_{j=1}^{i}\frac{1}{q_i} .$$

Proof. The overhead with frequencies $\boldsymbol{w}$ is

$$\max_{f\in M} \sum_i \frac{f(w_i)^2}{\|f(\boldsymbol{w})\|_1^2}\frac{1}{q_i} . \tag{7}$$

It suffices to consider f that are threshold functions. The expression for the threshold at w_i has $f(w_j)/\|f(\boldsymbol{w})\|_1 = 1/i$ for $j \leq i$ and 0 otherwise. We get that the sum is $\frac{1}{i^2}\sum_{j=1}^{i}\frac{1}{q_i}$. The claim follows from taking the maximum over all threshold functions. □

In our context, the probabilities $\boldsymbol{q}$ are not something we directly control but rather emerge as an artifact of applying a certain sampling scheme to a dataset with certain frequencies $\boldsymbol{w}$. We will explore the universal overhead of $\boldsymbol{q}$ we obtain when applying off-the-shelf schemes to Zipf frequencies and to our datasets.

For Zipf$[\alpha]$ frequencies ($\alpha \geq 1/2$), ℓ_p sampling with $p = 1/\alpha$ is a universal emulator with (optimal) overhead H_n. Interestingly, for $\alpha \geq 1/2$, this is attained by ℓ_p sampling with $p \leq 2$, which has polylogarithmic size sketches. Note that we match here a different ℓ_p sample for each possible Zipf parameter α of the data frequencies. A sampling scheme that emulates ℓ_p sampling for a range $[p_1, p_2]$ of p values with some overhead h will be a universal emulator with overhead hH_n for Zipf$[\alpha]$ for $\alpha \in [1/p_2, 1/p_1]$. One such sampling scheme with polylogarithmic-size sketches was provided in (Cohen, 2018; 2015). The sample emulates all concave sublinear functions that include capping functions $f(w) = \min\{w, T\}$ for $T > 0$ and low moments with $p \in [0, 1]$ with $O(\log n)$ overhead.

Table 3 reports the universal overhead on our datasets with ℓ_1, ℓ_2, and multi-objective concave-sublinear sampling probabilities. We observe that while ℓ_2 sampling emulates high moments extremely well, its universal overhead is very large due to poor emulation of "slow growth" functions. The better universal overhead of ℓ_1 and concave-sublinear samples is $h \in [143, 700]$ and is practically meaningful as it is in the regime where $h\varepsilon^{-2} \ll n$.

6. Conclusion

We propose a framework where performance and statistical guarantees of sampling sketches are analyzed in terms of supported *frequencies-functions combinations*. We demonstrate analytically and empirically that sketches originally designed to sample according to "easy" functions of frequency on "hard" frequency distributions turn out to be accurate for sampling according to "hard" functions of frequency on "practical" frequency distributions. In particular, on "practical" distributions we can accurately approximate high frequency moments ($p > 2$) and the rank versus frequency distribution using small composable sketches.

Acknowledgments

We are grateful to the authors of (Hsu et al., 2019), especially Chen-Yu Hsu and Ali Vakilian, for sharing their data, code, and predictions with us. We thank Ravi Kumar and Robert Krauthgamer for helpful discussions. The work of Rasmus Pagh was supported by Investigator Grant 16582, Basic Algorithms Research Copenhagen (BARC), from the VILLUM Foundation.

References

Aamand, A., Indyk, P., and Vakilian, A. (Learned) frequency estimation algorithms under zipfian distribution. *CoRR*, abs/1908.05198, 2019. URL `http://arxiv.org/abs/1908.05198`.

Agarwal, P. K., Cormode, G., Huang, Z., Phillips, J. M., Wei, Z., and Yi, K. Mergeable summaries. *ACM Trans. Database Syst.*, 38(4):26:1–26:28, 2013. doi: 10.1145/2500128. URL `https://doi.org/10.`

1145/2500128.

Alon, N., Matias, Y., and Szegedy, M. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137 – 147, 1999. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1545.

Alon, N., Gibbons, P. B., Matias, Y., and Szegedy, M. Tracking join and self-join sizes in limited storage. *Journal of Computer and System Sciences*, 64(3):719–747, 2002.

Andoni, A., Krauthgamer, R., and Onak, K. Streaming algorithms via precision sampling. In *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, 2011. URL https://doi.org/10.1109/FOCS.2011.82.

Braverman, V. and Ostrovsky, R. Zero-one frequency laws. In *STOC*. ACM, 2010.

CAIDA. The caida ucsd anonymized internet traces 2016 - 2016/01/21 13:29:00 utc. https://www.caida.org/data/passive/passive_dataset.xml, 2016.

Charikar, M., Chen, K., and Farach-Colton, M. Finding frequent items in data streams. In *Proceedings of the International Colloquium on Automata, Languages and Programming (ICALP)*, pp. 693–703, 2002.

Cohen, E. Multi-objective weighted sampling. In *2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb)*, pp. 13–18, 2015. doi: 10.1109/HotWeb.2015.8. Full version: https://arxiv.org/abs/1509.07445.

Cohen, E. Stream sampling framework and application for frequency cap statistics. *ACM Trans. Algorithms*, 14 (4):52:1–52:40, 2018. ISSN 1549-6325. doi: 10.1145/3234338.

Cohen, E. and Geri, O. Sampling sketches for concave sublinear functions of frequencies. In *NeurIPS*, 2019.

Cohen, E. and Kaplan, H. Tighter estimation using bottom-k sketches. In *Proceedings of the 34th VLDB Conference*, 2008. URL http://arxiv.org/abs/0802.3448.

Cohen, E., Kaplan, H., and Sen, S. Coordinated weighted sampling for estimating aggregates over multiple weight assignments. *VLDB*, 2, 2009. full version: http://arxiv.org/abs/0906.4560.

Cohen, E., Cormode, G., and Duffield, N. Don't let the negatives bring you down: Sampling from streams of signed updates. In *Proc. ACM SIGMETRICS/Performance*, 2012.

Cormode, G. and Muthukrishnan, S. An improved data stream summary: The count-min sketch and its applications. *J. Algorithms*, 55(1), 2005.

Duffield, N., Thorup, M., and Lund, C. Priority sampling for estimating arbitrary subset sums. *Journal of the ACM*, 54(6), 2007.

Eden, T., Ron, D., and Seshadhri, C. Sublinear time estimation of degree distribution moments: The arboricity connection. *SIAM J. Discrete Math.*, 33, 2019. URL https://doi.org/10.1137/17M1159014.

Estan, C. and Varghese, G. New directions in traffic measurement and accounting. In *SIGCOMM*. ACM, 2002.

Flajolet, P. and Martin, G. N. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31:182–209, 1985.

Flajolet, P., Fusy, E., Gandouet, O., and Meunier, F. Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm. In *Analysis of Algorithms (AofA)*. DMTCS, 2007.

Frieze, A., Kannan, R., and Vempala, S. Fast monte-carlo algorithms for finding low-rank approximations. *J. ACM*, 51(6), 2004. URL https://doi.org/10.1145/1039488.1039494.

Gibbons, P. and Matias, Y. New sampling-based summary statistics for improving approximate query answers. In *SIGMOD*. ACM, 1998.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459. URL http://www.jstor.org/stable/2280784.

Hsu, C.-Y., Indyk, P., Katabi, D., and Vakilian, A. Learning-based frequency estimation algorithms. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1lohoCqY7.

Indyk, P. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *Proc. 41st IEEE Annual Symposium on Foundations of Computer Science*, pp. 189–197. IEEE, 2001.

Indyk, P. and Woodruff, D. P. Optimal approximations of the frequency moments of data streams. In *STOC*, pp. 202–208. ACM, 2005.

Indyk, P., Vakilian, A., and Yuan, Y. Learning-based low-rank approximations. In *NeurIPS*, 2019.

Jayaram, R. and Woodruff, D. P. Perfect lp sampling in a data stream. In *FOCS*, 2018. URL https://doi.org/10.1109/FOCS.2018.00058.

Jiang, T., Li, Y., Lin, H., Ruan, Y., and Woodruff, D. P. Learning-augmented data stream algorithms. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HyxJ1xBYDH.

Kraska, T., Beutel, A., Chi, E. H., Dean, J., and Polyzotis, N. The case for learned index structures. In *SIGMOD*. Association for Computing Machinery, 2018. URL https://doi.org/10.1145/3183713.3196909.

Li, Y. and Woodruff, D. P. A tight lower bound for high frequency moment estimation with small error. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pp. 623–638. Springer, 2013.

Liu, Z., Manousis, A., Vorsanger, G., Sekar, V., and Braverman, V. One sketch to rule them all: Rethinking network flow monitoring with univmon. In *SIGCOMM*, 2016.

Liu, Z., Ben-Basat, R., Einziger, G., Kassner, Y., Braverman, V., Friedman, R., and Sekar, V. Nitrosketch: robust and general sketch-based monitoring in software switches. In *Proceedings of the ACM Special Interest Group on Data Communication, SIGCOMM 2019, Beijing, China, August 19-23, 2019*, pp. 334–350, 2019. URL https://doi.org/10.1145/3341302.3342076.

Maciá-Fernández, G., Camacho, J., Magán-Carrión, R., García-Teodoro, P., and Therón, R. UGR'16: A new dataset for the evaluation of cyclostationarity-based network idss. *Computers & Security*, 73:411 – 424, 2018. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2017.11.004. URL http://www.sciencedirect.com/science/article/pii/S0167404817302353.

Manku, G. and Motwani, R. Approximate frequency counts over data streams. In *International Conference on Very Large Databases (VLDB)*, pp. 346–357, 2002.

McGregor, A., Vorotnikova, S., and Vu, H. T. Better algorithms for counting triangles in data streams. In *Proceedings of the 35th Symposium on Principles of Database Systems (PODS)*, pp. 401–411, 2016.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Aguera y Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*. PMLR, 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

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

Misra, J. and Gries, D. Finding repeated elements. Technical report, Cornell University, 1982.

Ohlsson, E. Sequential poisson sampling. *J. Official Statistics*, 14(2):149–162, 1998.

Paranjape, A., Benson, A. R., and Leskovec, J. Motifs in temporal networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM 17, pp. 601–610, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346757. doi: 10.1145/3018661.3018731. URL https://doi.org/10.1145/3018661.3018731.

Pass, G., Chowdhury, A., and Torgeson, C. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, InfoScale 06, pp. 1–es, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595934286. doi: 10.1145/1146847.1146848. URL https://doi.org/10.1145/1146847.1146848.

Rosén, B. Asymptotic theory for successive sampling with varying probabilities without replacement, I. *The Annals of Mathematical Statistics*, 43(2):373–397, 1972. URL http://www.jstor.org/stable/2239977.

Rosén, B. Asymptotic theory for order sampling. *J. Statistical Planning and Inference*, 62(2):135–158, 1997.