Noname manuscript No.
(will be inserted by the editor)

Space-Efficient Estimation of Statistics over Sub-Sampled Streams

**Andrew McGregor · A. Pavan ·
Srikanta Tirthapura · David Woodruff**

the date of receipt and acceptance should be inserted later

Abstract In many stream monitoring situations, the data arrival rate is so high that it is not even possible to observe each element of the stream. The most common solution is to sub-sample the data stream and use the sample to infer properties and estimate aggregates of the original stream. However, in many cases, the estimation of aggregates on the original stream cannot be accomplished through simply estimating them on the sampled stream, followed by a normalization. We present algorithms for estimating frequency moments, support size, entropy, and heavy hitters of the original stream, through a single pass over the sampled stream.

Keywords data streams, frequency moments, sub-sampling

1 Introduction

In many stream monitoring situations, the data arrival rate is so high that it is possible to observe each element in the stream. The most common solution is to sub-sample the data stream and use the sample to infer properties of the original stream. For example, in an IP router, aggregated statistics of the packet stream are maintained through a protocol such as Netflow [9]. In high-end routers, the load due to statistics maintenance can be so high that a variant of Netflow called *sampled Netflow* has been developed. In randomly sampled netflow, the monitor gets to view only a random sample of the packet stream, and must maintain statistics on the original stream, using this view.

In such scenarios of extreme data deluge, we are faced with two constraints on data processing. First, the entire data set is not seen by the monitor; only a random sample is

Andrew McGregor
University of Massachusetts, E-mail: mcgregor@cs.umass.edu. Supported by NSF CAREER Award CCF-0953754.

A. Pavan
Iowa State University, E-mail: pavan@cs.iastate.edu. Supported in part by NSF CCF-0916797.

Srikanta Tirthapura
Iowa State University, E-mail: snt@iastate.edu. Supported in part by NSF CNS-0834743, CNS-0831903.

David P. Woodruff
IBM Almaden, E-mail: dpwoodru@us.ibm.com

seen. Second, even the random sample of the input is too large to be stored in main memory (or in secondary memory), and must be processed in a single pass through the data, as in the usual data stream model.

While there has been a large body of work that has dealt with data processing using a random sample (see for example, [3, 4]), and extensive work on the one-pass data stream model (see for example, [1, 29, 33]), there has been little work so far on data processing in the presence of both constraints, where only a random sample of the data set must be processed in a streaming fashion. We note that the estimation of frequency moments over a sampled stream is one of the open problems from [31], posed as Question 13, "Effects of Subsampling".

1.1 Problem Setting

We assume the setting of *Bernoulli sampling*, described as follows. Consider an input stream $P = \langle a_1, a_2, \ldots, a_n \rangle$ where $a_i \in \{1, 2, \ldots, m\}$. For a parameter p, $0 < p \leq 1$, a sub-stream of P, denoted L is constructed as follows. For $1 \leq i \leq n$, a_i is included in L with probability p. The stream processor is only allowed to see L, and cannot see P. The goal is to estimate properties of P through processing stream L. In the following discussion, L is called the *sampled stream*, and P is called the *original stream*.

1.2 Our Results

We present algorithms and lower bounds for estimating key aggregates of a data stream by processing a randomly sampled substream. We consider the basic frequency related aggregates, including the number of distinct elements, the frequency moments, the empirical entropy of the frequency distribution, and the heavy hitters.

1. **Frequency Moments:** For the frequency moments F_k for $k \geq 2$, we present $(1+\varepsilon, \delta)$-approximation algorithms with space complexity[1] $\tilde{O}(p^{-1} m^{1-2/k})$. This result yields an interesting tradeoff between the sampling probability and the space used by the algorithm. The smaller the sampling probability (up to a certain minimum probability), the greater is the streaming space complexity of our algorithm. The algorithm is presented in Section 3.
2. **Distinct Elements:** For the number of distinct elements, F_0, we show that the current best offline methods for estimating F_0 from a random sample can be implemented in a streaming fashion using very small space. While it is known that random sampling can significantly reduce the accuracy of an estimate for F_0 [7], we show that the need to process this stream using small space does not. The upper and lower bounds are presented in Section 4.
3. **Entropy:** For estimating entropy we first show that no multiplicative approximation is possible in general even when p is constant. However, we show that estimating the empirical entropy on the sampled stream yields a constant factor approximation to the entropy of the original stream if the entropy is larger than some vanishingly small function of p and n. These results are presented in Section 5.

[1] Where $\tilde{O}$ notation suppresses factors polynomial in $1/\varepsilon$ and $1/\delta$ and factors logarithmic in m and n.

4. **Heavy Hitters:** We show tight bounds for identifying a set of $O(1/\alpha)$ elements whose frequency exceeds $\alpha F_k^{1/k}$ for $k \in \{1,2\}$. In the case of $k = 1$, we show that existing heavy hitter algorithms can be used if the stream is sufficiently long compared with p. In the case of $k = 2$, we show how to adapt ideas used in Section 3 to arrive at an algorithm that uses space $\tilde{O}(1/p)$.

Another way of interpreting our results is in terms of time-space tradeoffs for data stream problems. Almost every streaming algorithm has a time complexity of at least n, since the algorithm reads and processes each stream update. We show that for estimating F_k (and other problems) it is unnecessary to process each update; instead, it suffices for the algorithm to read each item independently with probability p, and maintain a data structure of size $\tilde{O}(p^{-1} \cdot m^{1-2/k})$. Interestingly, the time to update the data structure per sampled stream item is still only $\tilde{O}(1)$. The time to output an estimate at the end of observation is $\tilde{O}(p^{-1} \cdot m^{1-2/k})$, i.e., roughly linear in the size of the data structure. As an example of the type of tradeoffs that are achievable, for estimating F_2 if $n = \Theta(m)$ we can set $p = \tilde{\Theta}(1/\sqrt{n})$ and obtain an algorithm using $\tilde{O}(\sqrt{n})$ total processing time and $\tilde{O}(\sqrt{n})$ workspace.

1.3 Related Work

There is a large body of prior work related at the intersection of random sampling and data stream processing. Some of this work is along the lines of methods for random sampling from a data stream, including the reservoir sampling algorithm, attributed to Waterman (also see [37]). There has been much follow up on variants and generalizations of reservoir sampling, see for example [2,16,20,30,36]. While this line of work focuses on how to efficiently sample from a stream, our work focuses on how to process a stream that has already been sampled.

Stream sampling is a well-researched method for managing the load on network monitors, while enabling accurate measurement. Packets are grouped into *flows* based on the values of certain attributes within the packet header. One commonly used sampling method is the "sampled netflow" model (NF) [23], which is the same as the Bernoulli sampling that we consider here, where packets are sampled independent of each other. Other methods of sampling are also considered under the general umbrella of sampled netflow, such as deterministic sampling (one of out every n packets). Another sampling method is the sample-and-hold model (SH) [22], where, once a packet is sampled from a flow, all other packets belonging that flow are also sampled. The priority sampling procedure [19] is a method for sampling from a weighted stream so that we can get unbiased estimators of individual weights with small variance. Szegedy [35] has shown that the priority sampling method of [19] essentially gets the smallest possible variance, given a fixed sample size. In addition, various combinations and enhancements to these sampling mechanisms have been proposed [10–12, 21]. In particular, [12] presents methods for better tuning sampling parameters and for exporting partial summaries to slower storage, [21] presents methods that dynamically adapt the sampling rate to achieve a desired level of accuracy, [10] present structure-aware sampling methods that provide improved accuracy (when compared with NF) on specific range queries of interest, and [11] presents stream sampling schemes for variance-optimal estimation of the total weight of an arbitrary subset of the stream of a certain size. There is much other work along the lines of optimizing sampling methods for accurate estimation of a specific class of aggregates on the original stream. Typical aggregates of interest include the distribution of the number of packets in different flows, and

aggregates over sub-populations of all flows. The above line of work tailors the sampling scheme towards specific goals, while we consider a simple but general sampling scheme, Bernoulli sampling, and explore how to efficiently process data under this sampling strategy. In many situations, including with sampled netflow, the sampling strategy is already decided by an external entity, such as the router, over which we may not have control.

Duffield et al. [17] consider the estimation of the sizes of IP flows and the number of IP flows in a packet stream through observing the sampled stream. In a follow up work [18], they provide methods for estimating the distribution of the sizes of the input flows by observing samples of the original stream; this can be viewed as constructing an approximate histogram. The techniques used here are maximum likelihood estimation, as well as protocol level detail at the IP and TCP level. Other work along this lines includes the work on inverting sampled traffic [26] which aims to recover the distribution of the original traffic through analyzing the sample, and work in [5, 13] which seeks to answer top-k queries and rank flows through analyzing the sample. While this line of work deals with inference from a random sample in detail, it does not consider the issue of processing the sample in a streaming manner using limited space, as we do here.

Further, we consider aggregates such as frequency moments and entropy, which do not seem to have been investigated in detail on sampled streams in prior work on network monitoring. In particular, even when the space complexity of an algorithm is high, we present space lower bounds that help understand the extend to which these aggregates can be estimated.

Rusu and Dobra [34] consider the estimation of the second frequency moment of a stream, equivalently, the size of the self-join, through processing the sampled stream. Our work differs from theirs in the following ways. While [34] do not explicitly mention the space bound of their algorithm, we derived an $(1+\varepsilon,\delta)$ estimator for F_2 based on their algorithm and found that the estimator took $\tilde{O}(1/p^2)$ space. We improve the dependence on the sampling probability and obtain an algorithm that only requires $\tilde{O}(1/p)$ space. This dependence on the sampling probability p is optimal. Our technique is also different from theirs. Ours relies on counting the number of collisions in the sampled stream, while theirs relies on scaling an estimate of the second frequency moment of the sampled stream. We also consider higher frequency moments F_k, for $k > 2$, as well as the entropy, while they do not.

Bhattacharya et al. [6] consider stream processing in the model where the stream processor can adaptively "skip" past stream elements, and look at only a fraction of the input stream, thus speeding up stream computation. In their model, the stream processor has the power to decide which elements to see and which to skip past, hence it is "adaptive"; in our model, the stream processor does not have such power, and must deal with the randomly sampled stream that is presented to it. Our model reflects the setup in current network monitoring equipment, such as Randomly Sampled Netflow [9]. They present a constant factor approximation for F_2, while we present $(1+\varepsilon,\delta)$ approximations for all frequency moments F_k for $k \geq 2$.

Bar-Yossef [3] presents lower bounds on the sampling probability, or equivalently, the number of samples needed to estimate certain properties of a data set, including the frequency moments. This yields a minimum sampling probability for the Bernoulli sampler that we consider, below which it is not possible to estimate aggregates accurately, whether streaming or otherwise. This is relevant to Theorem 1 in our paper, which assumes that the sampling probability must be at least a certain value.

There is work on *probabilistic data streams* [14,28], where the data stream itself consists of "probabilistic" data, and each element of the stream is a probability distribution over a

set of possible events. Unlike in our model, the stream processor gets to see the entire input in the probabilistic streams model.

Remark. The preliminary conference version of this paper claimed matching lower bounds for estimating F_k and heavy hitters [32]. The claimed lower bounds crucially depend on lower bounds obtained in an earlier work of Guha and Huang [24]. However, a problem has been found with the bounds of [24]. Thus the lower bound proofs that were presented in [32] do not hold.

2 Notation and Preliminaries

Throughout this paper, we will denote the original length-n stream by $P = \langle a_1, a_2, \ldots, a_n \rangle$ and will assume that each element $a_i \in \{1, 2, \ldots, m\}$. We denote the sampling probability with p. The sampled stream L is constructed by including each a_i in L with probability p, independent of the other elements. It is assumed that the sampling probability p is fixed in advance and is known to the algorithm.

Throughout let f_i be the frequency of item i in the original stream P. Let g_i be the frequency in the sub-sampled stream and note that $g_i \sim \mathrm{Bin}(f_i, p)$. The streams P and L define frequency vectors $\mathbf{f} = (f_1, f_2, \ldots, f_m)$ and $\mathbf{g} = (g_1, g_2, \ldots, g_m)$ respectively.

When considering a function F on a stream (e.g., a frequency moment or the entropy) we will denote $F(P)$ and $F(L)$ to indicate that value of the function on the original and sampled stream respectively. When the context is clear, we will also abuse notation and use F to indicate $F(P)$. We are primarily interested in randomized multiplicative approximations.

Definition 1 For $\alpha > 1$ and $\delta \in [0,1]$, we say $\tilde{X}$ is an an (α, δ)-estimator for X if

$$\Pr\left[\alpha^{-1} \leq X/\tilde{X} \leq \alpha\right] \geq 1 - \delta \ .$$

We use the notation $\tilde{O}$ to suppress factors polynomial in $1/\varepsilon$, $1/\delta$ and logarithmic in n. More precisely, given two functions f and g and constants $\varepsilon > 0$, and $\delta > 0$, we write $f(n) \in \tilde{O}(g(n))$ to denote $f(n) \in O((poly(1/\varepsilon, 1/\delta, \log n)g(n))$. Similarly we write $f(n) \in \tilde{\Omega}(g(n))$ to denote $f(n) \in \Omega((poly(1/\varepsilon, 1/\delta, \log n)g(n))$.

3 Frequency Moments

In this section, we present an algorithm for estimating the kth frequency moment F_k. The main theorem of this section is as follows.

Theorem 1 *For $k \geq 2$, there is a one pass streaming algorithm which observes L and outputs a $(1+\varepsilon, \delta)$-estimator for $F_k(P)$ using $\tilde{O}(p^{-1}m^{1-2/k})$ space, assuming $p = \tilde{\Omega}(\min(m,n)^{-1/k})$.*

For $p = \tilde{o}(\min(m,n)^{-1/k})$ there is not enough information in the sampled stream to obtain an $(1+\varepsilon, \delta)$ approximation to $F_k(P)$ with any amount of space, see Theorem 4.33 of [3].

Definition 2 For $1 \leq \ell \leq k$ define the number of ℓ-wise collisions to be $C_\ell(P) = \sum_{i=1}^{m} \binom{f_i}{\ell}$ and $C_\ell(L) = \sum_{i=1}^{m} \binom{g_i}{\ell}$.

Our algorithm is based on the following connection between the ℓth frequency moment of a stream and the ℓ-wise collisions in the stream.

Lemma 1 *For* $1 \leq \ell \leq k$,

$$F_\ell(P) = \ell! \cdot C_\ell(P) + \sum_{l=1}^{\ell-1} \beta_l^\ell F_l(P) \tag{1}$$

where $\beta_l^\ell = (-1)^{\ell-l+1} \sum_{1 \leq j_1 < \ldots < j_{\ell-l} \leq (\ell-1)} (j_1 \cdot j_2 \cdots j_{\ell-l})$.

Proof The relationship follows from

$$\begin{aligned}
\ell! \cdot C_\ell(P) &= \sum_{i=1}^{m} f_i(f_i-1)\ldots(f_i-(\ell-1)) \\
&= \sum_{i=1}^{m} \left(f_i^\ell - f_i^{\ell-1} \cdot \left(\sum_{1 \leq j_1 \leq \ell-1} j_1 \right) + f_i^{\ell-2} \cdot \left(\sum_{1 \leq j_1 < j_2 \leq \ell-1} j_1 \cdot j_2 \right) - \ldots \right) \\
&= \sum_{i=1}^{m} f_i^\ell - \left(\sum_{1 \leq j_1 \leq \ell-1} j_1 \right) \cdot \sum_{i=1}^{m} f_i^{\ell-1} + \left(\sum_{1 \leq j_1 < j_2 \leq \ell-1} j_1 \cdot j_2 \right) \cdot \sum_{i=1}^{m} f_i^{\ell-2} - \ldots \\
&= F_\ell(P) - \sum_{l=1}^{\ell-1} \beta_l^\ell F_l(P) \; .
\end{aligned}$$

□

The following lemma relates the expectation of $C_\ell(L)$ to $C_\ell(P)$ and bounds the variance.

Lemma 2 *For* $1 \leq \ell \leq k$, $\mathbb{E}[C_\ell(L)] = p^\ell C_\ell(P)$ *and* $\mathbb{V}[C_\ell(L)] = O(p^{2\ell-1} F_\ell^{2-1/\ell})$.

Proof Let C denote $C_\ell(L)$. Since each ℓ-wise collision in P appears in L with probability p^ℓ, we have $\mathbb{E}[C] = p^\ell C_\ell(P)$. For each $i \in [m]$, let C_i be the number of ℓ-wise collisions in L among items that equal i. Then $C = \sum_{i \in [m]} C_i$. By independence of the C_i,

$$\mathbb{V}[C] = \sum_{i \in [m]} \mathbb{V}[C_i] \; .$$

Fix an $i \in [m]$. Let S_i be the set of indices in the original stream equal to i. For each $J \subseteq S_i$ with $|J| = \ell$, let X_J be an indicator random variable if each of the stream elements in J appears in the sampled stream. Then $C_i = \sum_J X_J$. Hence,

$$\begin{aligned}
\mathbb{V}[C_i] &= \sum_{J,J'} \mathbb{E}[X_J X_{J'}] - \mathbb{E}[X_J]\,\mathbb{E}[X_{J'}] \\
&= \sum_{J,J'} p^{|J \cup J'|} - p^{2\ell} \\
&= \sum_{j=1}^{\ell} \binom{f_i}{j} \cdot \binom{f_i - j}{2\ell - 2j} \cdot \binom{2\ell - 2j}{\ell - j} \cdot (p^{2\ell-j} - p^{2\ell}) \\
&= \sum_{j=1}^{\ell} O(f_i^{2\ell-j} p^{2\ell-j}).
\end{aligned}$$

Since $F_{2\ell-j}^{1/(2\ell-j)} \leq F_\ell^{1/\ell}$ for all $j = 1,\ldots,\ell$, we have

$$\mathbb{V}[C] = O(1)\cdot\sum_{j=1}^{\ell} F_{2\ell-j}\cdot p^{2\ell-j} = O(1)\cdot\sum_{j=1}^{\ell} F_\ell^{2-j/\ell}\cdot p^{2\ell-j}.$$

If we can show that the first term of this sum dominates, the desired variance bound follows. This is the case if $p\cdot F_\ell^{1/\ell} \geq 1$, since this is the ratio of two consecutive summands. Note that F_ℓ is minimized for a fixed F_0 and F_1 when there are F_0 frequencies each of value F_1/F_0. In this case,

$$F_\ell^{1/\ell} = (F_0\cdot(F_1/F_0)^\ell)^{1/\ell} = F_1/F_0^{1-1/\ell}.$$

Hence, $p \geq 1/F_\ell^{1/\ell}$ if $p \geq F_0^{1-1/\ell}/F_1$, which holds by assumption. □

We next describe the intuition behind our algorithm. To estimate $F_k(P)$, by Eq. 1, it suffices to obtain estimates for $F_1(P)$, $F_2(P),\ldots,F_{k-1}(P)$ and $C_k(P)$ (one of the caveats is that some of the coefficients of $F_i(P)$ are negative, which we handle as explained below). Our algorithm attempts to estimate $F_\ell(P)$ for $\ell = 1,2,\ldots$ inductively. Since, by Chernoff bounds, $F_1(P)$ is very close to $F_1(L)/p$, $F_1(P)$ can be estimated easily. Thus our problem reduces to estimating $C_k(P)$ by observing the sub-sampled stream L. Since the expected number of collisions in L equals $p^kC_k(P)$, our algorithm will attempt to estimate $C_k(L)$, the number of k-wise collisions in the sub-sampled stream. However, it is not possible to find a good relative approximation of $C_k(L)$ in small space if $C_k(L)$ is small. However, when $C_k(L)$ is small, it does not contribute significantly to the final answer and we do not need a good relative error approximation! We only need that our estimator does not grossly over estimate $C_k(L)$. Our algorithm to estimate $C_k(L)$ will have the following property: If $C_k(L)$ is large, then it outputs a good relative error approximation, and if $C_k(L)$ is small the it outputs a value that is at most $3C_k(L)$. Another caveat is that some of the β_i^ℓ's could be negative. Thus apriori it is not clear that our strategy of estimating $F_\ell(P)$ by estimating $F_1(P)$, $F_2(P),\ldots,F_{k-1}(P)$, $C_k(P)$, and applying Equation 1 works. However, by using a careful choice of approximation errors and the fact that $F_i(P) \geq F_j(P)$, when $i > j$, we argue that this approach succeeds in obtaining a good approximation of $F_\ell(P)$.

3.1 The Algorithm

Define a sequence of random variables ϕ_ℓ:

$$\phi_1 = \frac{F_1(L)}{p}, \quad \text{and} \quad \phi_\ell = \frac{C_\ell(L)\ell!}{p^\ell} + \sum_{i=1}^{\ell-1}\beta_i^\ell\phi_i \quad \text{for } \ell > 1.$$

Algorithm 1 inductively computes an estimate $\tilde{\phi}_i$ for each ϕ_i. Note that if $C_\ell(L)/p^\ell$ takes its expected value of $C_\ell(P)$ and we could compute $C_\ell(L)$ exactly, then Eq. 1 implies that the algorithm would return $F_k(P)$ exactly. While this is excessively optimistic we will show that $C_\ell(L)/p^\ell$ is sufficiently close to $C_\ell(P)$ with high probability and that we can construct an estimate for $\tilde{C}_\ell(L)$ for $C_\ell(L)$ such that the final result returned is still a $(1+\varepsilon)$ approximation for $F_k(P)$ with probability at least $1-\delta$.

Algorithm 1: $F_k(P)$

Compute $F_1(L)$ exactly and set $\tilde{\phi}_1 = F_1(L)/p$.
for $\ell = 2$ *to* k **do**
  Let $\tilde{C}_\ell(L)$ be an estimate for $C_\ell(L)$, computed as described in the text.
  Compute

$$\tilde{\phi}_\ell = \frac{\tilde{C}_\ell(L)\ell!}{p^\ell} + \sum_{i=1}^{\ell-1} \beta_i^\ell \tilde{\phi}_i$$

end
Return $\tilde{\phi}_k$.

We compute our estimate of $\tilde{C}_\ell(L)$ via an algorithm by Indyk and Woodruff [27]. This algorithm attempts to obtain a $1+\varepsilon_{\ell-1}$ approximation of $C_\ell(L)$ for some value of $\varepsilon_{\ell-1}$ to be determined. The estimator is as follows. For $i = 0, 1, 2, \ldots$ define

$$S_i = \{j \in [m] : \eta(1+\varepsilon')^i \le g_j < \eta(1+\varepsilon')^{i+1}\}$$

where η is randomly chosen between 0 and 1 and $\varepsilon' = \varepsilon_{\ell-1}/4$. The algorithm of Indyk and Woodruff [27] returns an estimate $\tilde{s}_i$ for $|S_i|$ and our estimate for $C_\ell(L)$ is defined as

$$\tilde{C}_\ell(L) := \sum_i \tilde{s}_i \binom{\eta(1+\varepsilon')^i}{\ell}$$

The space used by the algorithm is $\tilde{O}(p^{-1}m^{1-2/\ell})$. We defer the details to Section 3.2.

We next define an event $\mathscr{E}$ that corresponds to our collision estimates being sufficiently accurate and the sampled stream being "well-behaved." The next lemma establishes that $\Pr[\mathscr{E}] \ge 1-\delta$. We will defer the proof until Section 3.2.

Lemma 3 *Define the event* $\mathscr{E} = \mathscr{E}_1 \cap \mathscr{E}_2 \cap \ldots \cap \mathscr{E}_k$ *where*

$$\mathscr{E}_1 \;:\; \tilde{\phi}_1 \in (1 \pm \varepsilon_1) F_1(P)$$
$$\mathscr{E}_\ell \;:\; |\tilde{C}_\ell(L)/p^\ell - C_\ell(P)| \le \varepsilon_{\ell-1} F_\ell(P)/\ell! \quad \textit{for} \quad \ell \ge 2$$

where $\varepsilon_k = \varepsilon$, $\varepsilon_{\ell_1} = \frac{\varepsilon_\ell}{(A_\ell+1)}$, *and* $A_\ell = \sum_{i=1}^{\ell-1} |\beta_i^\ell|$. *Then* $\Pr[\mathscr{E}] \ge 1-\delta$.

The next theorem establishes that, conditioned on the event $\mathscr{E}$, the algorithm returns a $(1 \pm \varepsilon)$ approximation of $F_k(P)$ as required.

Lemma 4 *Conditioned on* $\mathscr{E}$, *we have* $\tilde{\phi}_\ell \in (1 \pm \varepsilon_\ell) F_\ell(P)$ *for all* $\ell \in [k]$.

Proof The proof is by induction on ℓ. Since we are conditioning on event $\mathscr{E}$ (and thus event $\mathscr{E}_1$), we have that $\tilde{\phi}_1$ is an $(1 \pm \varepsilon_1)$ approximation of $F_1(P)$. Thus the induction hypothesis ensures that $\tilde{\phi}_i$, $1 \le i \le \ell-1$, is a $(1 \pm \varepsilon_i)$ approximation of $F_i(P)$. Therefore,

$$\begin{aligned}
|\tilde{\phi}_\ell - F_\ell(P)| &= \left| \frac{\tilde{C}_\ell(L)\ell!}{p^\ell} + \sum_{i=1}^{\ell-1} \beta_i^\ell \tilde{\phi}_i - F_\ell(P) \right| \\
&\le \left| \ell! C_\ell(P) + \sum_{i=1}^{\ell-1} \beta_i^\ell F_i(P) - F_\ell(P) \right| + \varepsilon_{\ell-1} F_\ell(P) + \sum_{i=1}^{\ell-1} |\beta_i^\ell| F_i(P) \\
&= \varepsilon_{\ell-1} F_\ell(P) + \sum_{i=1}^{\ell-1} |\beta_i^\ell| \varepsilon_i F_i(P)
\end{aligned}$$

where the first inequality follows since we are conditioning on event $\mathscr{E}_\ell$ which ensures that

$$\left|\frac{\tilde{C}_\ell(L)\ell!}{p^\ell} - \ell! C_\ell(P)\right| \leq \varepsilon_{\ell-1} F_\ell(P),$$

and the induction hypothesis ensures that

$$\left|\sum_{i=1}^{\ell-1} \beta_i^\ell \tilde{\phi}_i - \sum_{i=1}^{\ell-1} \beta_i^\ell F_i(P)\right| \leq \sum_{i=1}^{\ell-1} |\beta_i^\ell| \varepsilon_i F_i(P) \ .$$

The second equality follows due to Equation 1. Note that $i \leq j$ implies $\varepsilon_i \leq \varepsilon_j$ and $F_i(P) \leq F_j(P)$. Hence, by the definition of ε_ℓ,

$$\varepsilon_{\ell-1} F_\ell(P) + \sum_{i=1}^{\ell-1} |\beta_i^\ell| \varepsilon_i F_i(P) \leq \varepsilon_{\ell-1} F_\ell(P) \left(1 + \sum_{i=1}^{\ell-1} |\beta_i^\ell|\right) = \varepsilon_\ell F_\ell(P) \ .$$

Therefore $\tilde{\phi}_\ell \in (1 \pm \varepsilon_\ell) F_\ell(P)$ as required. □

3.2 Proof of Lemma 3.

Our goal is to show that $\Pr[\mathscr{E}_1 \cap \mathscr{E}_2 \cap \ldots \cap \mathscr{E}_k] \geq 1 - \delta$. To do this it will suffice to show that for each $\ell \in [k]$, $\Pr[\mathscr{E}_\ell] \geq 1 - \delta/k$ and appeal to the union bound.

We first observe that, by Chernoff bounds, the event $\mathscr{E}_1$ happens with probability at least $1 - \delta/k$. Let X_i denote the 0-1 random variable whose value if 1 if the i item of the original stream appears in the sampled stream. Note that $E[X_i] = 1$, $1 \leq i \leq n$, and $F_1(L) = \sum_{i=1}^n X_i$. Since $\tilde{\phi}_1 = F_1(L)/p$, we have $\tilde{\phi}_1 = \sum_{i=1}^n X_i/p$. Recall that $n = F_1(P)$.

$$\begin{aligned}
\Pr\left[\overline{\mathscr{E}_1}\right] &= \Pr\left[|\tilde{\phi}_1 - F_1(P)| \geq F_1(P)\varepsilon_1\right] \\
&= \Pr\left[|\frac{\sum X_i}{p} - F_1(P)| \geq F_1(P)\varepsilon_1\right] \\
&= \Pr\left[|\frac{\sum X_i}{F_1(P)} - p \geq p\varepsilon_1\right] \\
&\leq 2e^{-\varepsilon_1^2 F_1(P) p/2} (\text{ By Chernoff Bound}) \\
&\leq \delta/k
\end{aligned}$$

The last inequality follows because our condition on p implies $p > \frac{poly(1/\varepsilon)\log 1/\delta}{F_1(p)}$.

To analyze $\Pr[\mathscr{E}_\ell]$ for $2 \leq \ell \leq k$ we consider the events:

$$\begin{aligned}
\mathscr{E}_\ell^1 &: \left|C_\ell(L)/p^\ell - C_\ell(P)\right| \leq \frac{\varepsilon_{\ell-1} F_\ell(P)}{2\ell!} \\
\mathscr{E}_\ell^2 &: \left|\tilde{C}_\ell(L)/p^\ell - C_\ell(L)/p^\ell\right| \leq \frac{\varepsilon_{\ell-1} F_\ell(P)}{2\ell!} \ .
\end{aligned}$$

By the triangle inequality it is easy to see that $\Pr\left[\mathscr{E}_\ell^1 \cap \mathscr{E}_\ell^2\right] \leq \Pr[\mathscr{E}_\ell]$ and hence it suffices to show that $\Pr\left[\mathscr{E}_\ell^1\right] \geq 1 - \delta/(2k)$ and $\Pr\left[\mathscr{E}_\ell^2\right] \geq 1 - \delta/(2k)$. The first part follows easily from the variance bound in Lemma 2.

Lemma 5 $\Pr\left[\mathscr{E}_\ell^1\right] \geq 1 - \frac{\delta}{4k}$.

Proof There are two cases depending on the value of $\mathbb{E}[C_\ell(L)]$.

Case I: First assume $\mathbb{E}[C_\ell(L)] \leq \frac{\delta\varepsilon_{\ell-1}p^\ell F_\ell}{8k\ell!}$. Therefore, by Lemma 2, we also know that

$$C_\ell(P) \leq \frac{\delta\varepsilon_{\ell-1}F_\ell}{8k\ell!} \ . \tag{2}$$

By Markov's bound

$$\Pr\left[C_\ell(L) \leq \frac{\varepsilon_{\ell-1}p^\ell F_\ell}{2\ell!}\right] \geq 1 - \frac{\delta}{4k} \ . \tag{3}$$

Eq. 2 and Eq. 3 together imply that with probability at least $1-\frac{\delta}{4k}$

$$\left|C_\ell(L)/p^\ell - C_\ell(P)\right| \leq \max\left(C_\ell(L)/p^\ell, C_\ell(P)\right) \leq \frac{\varepsilon_{\ell-1}F_\ell}{2\ell!}$$

Case II: Next assume $\mathbb{E}[C_\ell(L)] > \frac{\delta\varepsilon_{\ell-1}p^\ell F_\ell}{8k\ell!}$. By Chebyshev's bound, and using Lemma 2, we get:

$$\begin{aligned}
\Pr\left[|C_\ell(L) - \mathbb{E}[C_\ell(L)]| \geq \frac{\varepsilon_{\ell-1}\mathbb{E}[C_\ell(L)]}{2}\right] &\leq \frac{4\mathbb{V}[C_\ell(L)]}{\varepsilon_{\ell-1}^2(\mathbb{E}[C_\ell(L)])^2} \\
&\leq \frac{Dk^2(\ell!)^2}{\delta^2\varepsilon_{\ell-1}^4 pF_\ell^{1/\ell}} \\
&\leq \frac{Dk^2(\ell!)^2}{\delta^2\varepsilon_{\ell-1}^4 p}\frac{F_0^{1-1/\ell}}{F_1} \\
&\leq \frac{Dk^2(\ell!)^2}{\delta^2\varepsilon_{\ell-1}^4 p}\frac{1}{\min(F_0^{1/l}, F_1^{1/l})} \\
&= \frac{Dk^2(\ell!)^2}{\delta^2 H^4\varepsilon^4 p}\frac{1}{\min(F_0^{1/l}, F_1^{1/l})} \leq \frac{\delta}{4k}
\end{aligned}$$

where D and H are sufficiently large constants. The third inequality follows because $F_\ell^{1/\ell} \geq F_1/F_0^{1-1/\ell}$. The equality follows because $\varepsilon = H \times \varepsilon_{\ell-1}$. The last inequality follows because our assumption on p implies that $p \geq poly(1/\varepsilon, 1/\delta)\min(F_0, F_1)^{-1/k}$.

Since $\mathbb{E}[C_\ell(L)] = p^\ell C_\ell(P)$ and $C_\ell(P) \leq F_\ell(P)/\ell!$, we have that

$$\Pr\left[\left|C_\ell(L)/p^\ell - C_\ell(P)\right| \leq \frac{\varepsilon_{\ell-1}F_\ell(P)}{2\ell!}\right] \geq 1 - \frac{\delta}{4k}$$

as required. □

We will now show that $\mathscr{E}_\ell^2$ happens with high probability by analyzing the algorithm that computes $\tilde{C}_\ell(L)$. We need the following result due to Indyk and Woodruff [27]. Recall that $\varepsilon' = \varepsilon_{\ell-1}/4$.

Theorem 2 (Indyk and Woodruff [27]) *Let G be the set of indices i for which*

$$|S_i|(1+\varepsilon')^{2i} \geq \frac{\gamma F_2(L)}{\text{poly}(\varepsilon'^{-1}\log n)}, \tag{4}$$

then

$$\Pr\left[\forall i \in G,\ \tilde{s}_i \in (1\pm\varepsilon')|S_i|\right] \geq 1-\frac{\delta}{8k}.$$

For every i (whether it is in G or not) $\tilde{s}_i \leq 3|S_i|$. Moreover, the algorithm runs in space $\tilde{O}(1/\gamma)$.

We say that a set S_i contributes if

$$|S_i|\cdot\binom{(1+\varepsilon')^i}{\ell} > \frac{C_\ell(L)}{B}.$$

where $B=\mathrm{poly}(\varepsilon'^{-1}\log n)$. Given i the event that S_i contributes holds with certain (conceivably 0) probability. We first show that if S_i contributes, then S_i is a good set with high probability. More precisely, we show that for every S_i that contributes, Eq. (4) holds with high probability with $\gamma = pm^{-1+2/\ell}$.

Lemma 6 *Suppose that $C_\ell(L) > \frac{\varepsilon_{\ell-1}p^\ell F_\ell(P)}{4\ell!}$, and also suppose that the event S_i contributes happened. Then*

$$\Pr\left[|S_i|(1+\varepsilon')^{2i} \geq \frac{\delta p F_2(L)}{m^{1-2/\ell}\,\mathrm{poly}(\varepsilon'^{-1}\log n)}\right] \geq 1-\frac{\delta}{8k}.$$

Proof Consider a set S_i that contributes. Note that the probability that $\eta < 1/\mathrm{poly}(\delta^{-1}\varepsilon'^{-1}\log n)$ with is at most $1/\mathrm{poly}(\delta^{-1}\varepsilon'^{-1}\log n)$. Without loss of generality we can take this probability to be less than $\delta/16k$. By our assumption on $C_\ell(L)$ and the fact that S_i contributes,

$$|S_i|(1+\varepsilon')^{\ell i} \geq \frac{\varepsilon' p^\ell F_\ell(P)}{B\ell!}$$

holds with probability at least $1-\delta/8k$. Thus

$$|S_i|(1+\varepsilon')^{2i} \geq \frac{\varepsilon'^{2/\ell}p^2F_\ell^{2/\ell}(P)}{(B\ell!)^{2/\ell}} \geq \frac{p^2F_2(P)}{m^{1-2/\ell}\,\mathrm{poly}(\varepsilon'^{-1}\log n)}$$

where the second inequality is an application of Hölder's inequality.

Note that

$$\mathbb{E}\left[F_2(L)\right] = p^2F_2(P)+p(1-p)F_1(P) \leq pF_2(P)\ .$$

Thus, an application of the Markov bound,

$$\Pr\left[F_2(L) \leq \frac{16kpF_2(P)}{\delta}\right] \geq 1-\frac{\delta}{16k}. \tag{5}$$

The lemma follows as the following inequalities hold with probability at least $1-\delta/8k$.

$$\begin{aligned}
|S_i|(1+\varepsilon')^{2i} &\geq \frac{p^2F_2(P)}{m^{1-2/\ell}\,\mathrm{poly}(\varepsilon'^{-1}\log n)}\\
&\geq \frac{\delta p 16kpF_2(P)}{16km^{1-2/\ell}\,\mathrm{poly}(\varepsilon'^{-1}\log n)}\\
&\geq \frac{\delta pF_2(L)}{m^{1-2/\ell}\,\mathrm{poly}(\varepsilon'^{-1}\log n)} \quad \text{(By 5)}
\end{aligned}$$

□

Now we are ready to prove that the event $\mathscr{E}_\ell^2$ holds with high probability.

Lemma 7 $\Pr\left[\mathscr{E}_\ell^2\right] \geq 1 - \frac{\delta}{2k}$

Proof There are two cases depending on the size of $C_\ell(L)$.

Case I: Assume $C_\ell(L) \leq \frac{\varepsilon_{\ell-1} p^\ell F_\ell(P)}{4\ell!}$. By Theorem 2, it follows that $\tilde{C}_\ell(L) \leq 3C_\ell(L)$. Thus

$$\left|\tilde{C}_\ell(L) - C_\ell(L)\right| \leq 2C_\ell(L) \leq \frac{\varepsilon_{\ell-1} p^\ell F_\ell(P)}{2\ell!}$$

Case 2: Assume $C_\ell(L) > \frac{\varepsilon_{\ell-1} p^\ell F_\ell}{4\ell!}$. By Lemma 6, for every S_i that contributes,

$$\Pr\left[|S_i|(1+\varepsilon')^{2i} \geq \frac{\delta p F_2(L)}{m^{1-2/\ell}\,\mathrm{poly}(\varepsilon'^{-1}\log n)}\right] \geq 1 - \frac{\delta}{8k}.$$

Now by Theorem 2 for each S_i that contributes $\tilde{s}_i \in (1 \pm \varepsilon')|S_i|$, with probability at least $1 - \frac{\delta}{8k}$. Therefore,

$$\Pr\left[\left|\tilde{C}_\ell(L) - C_\ell(L)\right| \leq \varepsilon' C_\ell(L)\right] \geq 1 - \frac{\delta}{4k}\,.$$

If $\mathscr{E}_\ell^1$ is true, then:

$$C_\ell(L) \in C_\ell(P)p^\ell \pm \frac{\varepsilon_{\ell-1} F_\ell(P) p^\ell}{2\ell!}.$$

Since $\mathscr{E}_\ell^1$ holds with probability at least $1 - \frac{\delta}{4k}$, the following inequalities hold with probability at least $1 - \frac{\delta}{2k}$.

$$\begin{aligned}
\left|\tilde{C}_\ell(L) - C_\ell(L)\right| &\leq \varepsilon' C_\ell(L) \leq \varepsilon' C_\ell(P)p^\ell + \frac{\varepsilon_{\ell-1}\varepsilon' F_\ell(P)p^\ell}{2\ell!} \\
&\leq \frac{\varepsilon' F_\ell(P)p^\ell}{\ell!} + \frac{\varepsilon_{\ell-1}\varepsilon' F_\ell(P)p^\ell}{2\ell!} \\
&\leq \frac{F_\ell(P)p^\ell}{4\ell!}(\varepsilon_{\ell-1} + \varepsilon_{\ell-1}\varepsilon_{\ell-1}) \\
&\leq \frac{F_\ell(P)p^\ell \varepsilon_{\ell-1}}{2\ell!}
\end{aligned}$$

□

4 Distinct Elements

There are strong lower bounds for the accuracy of estimating the number of distinct values through random sampling. The following theorem is from Charikar et al. [7], which we have restated slightly to fit our notation (the original theorem is about database tables). Let F_0 be the number of elements in a data set T of total size n. Note that T maybe a stored data set, and need not be processed in a one-pass streaming manner.

Theorem 3 (Charikar et al. [7]) *Consider any (randomized) estimator $\hat{F}_0$ for the number of distinct values F_0 of T, that examines at most r out of the n elements in T. For any $\gamma > e^{-r}$, there exists a choice of the input T such that with probability at least γ, the multiplicative error is at least $\sqrt{(n-r)/(2r)\ln\gamma^{-1}}$.*

The above theorem implies that if we observe $o(n)$ elements of P, then it is not possible to get even an estimate with a constant multiplicative error. This lower bound for the non-streaming model leads to the following lower bound for sampled streams.

Theorem 4 **(F_0 Lower Bound)** *For sampling probability $p \in (0, 1/12]$, any algorithm that estimates F_0 by observing L, there is an input stream such that the algorithm will have a multiplicative error of $\Omega\left(1/\sqrt{p}\right)$ with probability at least $(1-e^{-np})/2$.*

Proof Let $\mathscr{E}_1$ denote the event $|L| \leq 6np$. Let β denote the multiplicative error of any algorithm (perhaps non-streaming) that estimates $F_0(P)$ by observing L. Let $\alpha = \sqrt{\frac{\ln 2}{12p}}$. Let $\mathscr{E}_2$ denote the event $\beta \geq \alpha$.

Note that $|L|$ is a binomial random variable. The expected size of the sampled stream is $\mathbb{E}[|L|] = np$. By using a Chernoff bound:

$$\Pr[\mathscr{E}_1] = 1 - \Pr[|L| > 6\mathbb{E}[|L|]] \geq 1 - 2^{-6\mathbb{E}[|L|]} > 1 - e^{-np}$$

If $\mathscr{E}_1$ is true, then the number of elements in the sampled stream is no more than $6np$. Substituting $r = 6np$ and $\gamma = 1/2$ in Theorem 3, we get:

$$\Pr[\mathscr{E}_2|\mathscr{E}_1] \geq \Pr\left[\beta > \sqrt{\left(\frac{n-6np}{12np}\right)\ln 2} \;\middle|\; \mathscr{E}_1\right] \geq \frac{1}{2}$$

Simplifying, and using $p \leq 1/12$, we get:

$$\Pr[\mathscr{E}_2] \geq \Pr[\mathscr{E}_1 \wedge \mathscr{E}_2] = \Pr[\mathscr{E}_1] \cdot \Pr[\mathscr{E}_2|\mathscr{E}_1] \geq \frac{1}{2}(1-e^{-np})$$

□

We now describe a simple streaming algorithm for estimating $F_0(P)$ by observing $L(P,p)$, which has an error of $O(1/\sqrt{p})$ with high probability.

Algorithm 2: $F_0(P)$

1 Let X denote a $(1/2, \delta)$-estimate of $F_0(L)$, derived using any streaming algorithm for F_0 (such as [29]).
2 Return $X/\sqrt{p}$

Lemma 8 **(F_0 Upper Bound)** *Algorithm 2 returns an estimate Y for $F_0(P)$ such that the multiplicative error of Y is no more than $4/\sqrt{p}$ with probability at least $1-(\delta + e^{-pF_0(P)/8})$.*

Proof Let $D = F_0(P)$, and $D_L = F_0(L)$. Let $\mathscr{E}_1$ denote the event $(D_L \geq pD/2)$, $\mathscr{E}_2$ denote $(X \geq D_L/2)$, and $\mathscr{E}_3$ denote the event $(X \leq 3D_L/2)$. Let $\mathscr{E} = \cap_{i=1}^{3}\mathscr{E}_i$.

Without loss of generality, let $1, 2, \ldots, D$ denote the distinct items that occurred in stream P. Define $X_i = 1$ if at least one copy of item i appeared in L, and 0 otherwise. The different X_is are all independent. Thus $D_L = \sum_{i=1}^{D} X_i$ is a the sum of independent Bernoulli random variables and

$$\mathbb{E}[D_L] = \sum_{i=1}^{D} \Pr[X_i = 1] \ .$$

Since each copy of item i is included in D_L with probability p, we have $\Pr[X_i = 1] \geq p$. Thus, $\mathbb{E}[D_L] \geq pD$. Applying a Chernoff bound,

$$\Pr\left[\overline{\mathscr{E}_1}\right] = \Pr\left[D_L < \frac{pD}{2}\right] \leq \Pr\left[D_L < \frac{\mathbb{E}[D_L]}{2}\right] \leq e^{-\mathbb{E}[D_L]/8} \leq e^{-pD/8}\,. \tag{6}$$

Suppose $\mathscr{E}$ is true. Then we have the following:

$$\frac{pD}{4} \leq \frac{D_L}{2} \leq X \leq \frac{3D_L}{2} \leq \frac{3D}{2}$$

The last inequality is because D_L is at most D. Therefore $X/\sqrt{p}$ has a multiplicative error of no more than $4/\sqrt{p}$.

We now bound the probability that $\mathscr{E}$ is false.

$$\Pr\left[\overline{\mathscr{E}}\right] \leq \sum_{i=1}^{3} \Pr\left[\overline{\mathscr{E}_i}\right] \leq \delta + e^{-pD/8}$$

where we have used the union bound, Eq. (6), and the fact that X is a $(1/2,\delta)$-estimator of D_L. □

5 Entropy

In this section we consider approximating the entropy of a stream.

Definition 3 The *entropy* of a frequency vector

$$\mathbf{f} = (f_1, f_2, \ldots, f_m)$$

is defined as $H(\mathbf{f}) = \sum_{i=1}^{m} \frac{f_i}{n} \lg \frac{n}{f_i}$ where $n = \sum_{i=1}^{m} f_i$.

Unfortunately, in contrast to F_0 and F_k, it is not possible to multiplicatively approximate $H(\mathbf{f})$ even if p is constant.

Lemma 9 *No multiplicative error approximation is possible with probability* $9/10$ *even with* $p > 1/2$ *. Furthermore,*

1. *There exists* $\mathbf{f}$ *such that* $H(\mathbf{f}) = \Theta(\log n/pn)$ *but* $H(\mathbf{g}) = 0$ *with probability at least* $9/10$*.*
2. *There exists* $\mathbf{f}$ *such that* $|H(\mathbf{f}) - H(\mathbf{g})| \geq |\lg(2p)|$ *with probability at least* $9/10$*.*

Proof First consider the following two scenarios for the contents of the stream. In Scenario 1, $f_1 = n$ and in Scenario 2, $f_1 = n-k$ and $f_2 = f_3 = \ldots = f_{k+1} = 1$. In the first case the entropy $H(\mathbf{f}) = 0$ whereas in the second,

$$\begin{aligned} H(\mathbf{f}) &= \frac{n-k}{n}(\lg e)\ln\frac{n}{n-k} + \frac{k}{n}\lg n \\ &= \frac{n-k}{n}\Theta(k/(n-k)) + \frac{k}{n}\lg n \\ &= (\Theta(1) + \lg n)\frac{k}{n}\,. \end{aligned}$$

Distinguishing these streams requires that at least one value other that 1 is present in the subsampled stream. This happens with probability $(1-p)^k > 1-pk$ and hence with $k = p^{-1}/10$ this probability is less than $9/10$.

For the second part of the lemma consider the stream with $f_1 = f_2 = \ldots = f_m = 1$ and hence $H(\mathbf{f}) = \lg m$. But $H(\mathbf{g}) = \lg |L|$ where $|L|$ is the number of elements in the sampled stream. By an application of the Chernoff bound $|L|$ is at most $2pm$ with probability at least $9/10$ and the result follows. □

Instead we will show that it is possible to approximate $H(\mathbf{f})$ up to a constant factor with an additional additive error term that tends to zero if $p = \omega(n^{-1/3})$. It will also be convenient to consider the following quantity:

$$H_{pn}(\mathbf{g}) = \sum_{i=1}^{m} \frac{g_i}{pn} \lg \frac{pn}{g_i} \ .$$

The following propositions establishes that $H_{pn}(\mathbf{g})$ is a very good approximation to $H(\mathbf{g})$.

Proposition 1 *With probability* $199/200$, $|H_{pn}(\mathbf{g}) - H(\mathbf{g})| = O(\log m/\sqrt{pn})$.

Proof By an application of the Chernoff bound, with probability $199/200$

$$|pn - \sum_{i=1}^{m} g_i| \leq c\sqrt{pn}$$

for some constant $c > 0$. Hence, if $n' = \sum_{i=1}^{m} g_i$ and $\gamma = n'/pn$ it follows that $\gamma = 1 \pm O(1/\sqrt{pn})$. Then

$$H_{pn}(\mathbf{g}) = \sum_{i=1}^{m} \frac{g_i}{pn} \lg \frac{pn}{g_i} = \sum_{i=1}^{m} \frac{\gamma g_i}{n'} \lg \frac{n'}{\gamma g_i} = H(\mathbf{g}) + O(1/\sqrt{pn}) + O(H(\mathbf{g})/\sqrt{pn}) \ .$$

□

The next lemma establishes that the entropy of $\mathbf{g}$ is within a constant factor of the entropy of $\mathbf{f}$ plus a small additive term.

Lemma 10 *With probability* $99/100$, *if* $p = \omega(n^{-1/3})$,

1. $H_{pn}(\mathbf{g}) \leq O(H(\mathbf{f}))$.
2. $H_{pn}(\mathbf{g}) \geq H(\mathbf{f})/2 - O\left(\frac{1}{p^{1/2}n^{1/6}}\right)$

Proof For the first part of the lemma, first note that

$$\mathbb{E}[H_{pn}(\mathbf{g})] = \sum_{i=1}^{m} \mathbb{E}\left[\frac{g_i}{pn} \lg \frac{pn}{g_i}\right] \leq \sum_{i=1}^{m} \frac{\mathbb{E}[g_i]}{pn} \lg \frac{pn}{\mathbb{E}[g_i]} = \sum_{i=1}^{m} \frac{pf_i}{pn} \lg \frac{pn}{pf_i} = H(\mathbf{f})$$

where the inequality follows from Jensen's inequality since the function $x \lg x^{-1}$ is concave. Hence, by Markov's inequality

$$\Pr[H_{pn}(\mathbf{g}) \leq 100H(\mathbf{f})] \geq 99/100 \ .$$

To prove the second part of the lemma, define $f^* = cp^{-1}\varepsilon^{-2}\log n$ for some sufficiently large constant c and $\varepsilon \in (0,1)$. We then partition $[m]$ into $A = \{i : f_i < f^*\}$ and $B = \{i : f_i \geq f^*\}$ and consider $H(\mathbf{f}) = H^A(\mathbf{f}) + H^B(\mathbf{f})$ where

$$H^A(\mathbf{f}) = \sum_{i\in A} \frac{f_i}{n} \lg \frac{n}{f_i} \quad \text{and} \quad H^B(\mathbf{f}) = \sum_{i\in B} \frac{f_i}{n} \lg \frac{n}{f_i} \ .$$

By applications of the Chernoff and union bounds, with probability at least $299/300$,

$$|g_i - pf_i| \leq \begin{cases} \varepsilon p f^* & \text{if } i \in A \\ \varepsilon p f_i & \text{if } i \in B \end{cases} .$$

Hence,

$$H^B_{pn}(\mathbf{g}) = \sum_{i\in B} \frac{g_i}{pn} \lg \frac{pn}{g_i} = \sum_{i\in B} \frac{f_i(1\pm\varepsilon)}{n} \lg \frac{n}{(1\pm\varepsilon)f_i} = (1\pm\varepsilon)H^B(\mathbf{f}) + O(\varepsilon) \ .$$

For $H^A_{pn}(\mathbf{g})$ we have two cases depending on whether $\sum_{i\in A} f_i$ is smaller or larger than $\theta := cp^{-1}\varepsilon^{-2}$. If $\sum_{i\in A} f_i \leq \theta$ then

$$H^A(\mathbf{f}) = \sum_{i\in A} \frac{f_i}{n} \lg \frac{n}{f_i} \leq \frac{\theta \lg n}{n} \ .$$

On the other hand if $\sum_{i\in A} f_i \geq \theta$ then by an application of the Chernoff bound,

$$|\sum_{i\in A} g_i - p\sum_{i\in A} f_i| \leq \varepsilon p \sum_{i\in A} f_i$$

and hence

$$\begin{aligned} H^A_{pn}(\mathbf{g}) &= \sum_{i\in A} \frac{g_i}{pn} \lg \frac{pn}{g_i} \\ &\geq \lg \frac{n}{(1+\varepsilon)f^*} \sum_{i\in A} \frac{g_i}{pn} \\ &\geq (1-\varepsilon) \lg \frac{n}{(1+\varepsilon)f^*} \sum_{i\in A} \frac{f_i}{n} \\ &\geq \left(1 - \varepsilon - \frac{\lg(1+\varepsilon)f^*}{\lg n}\right) H^A(\mathbf{f}) \ . \end{aligned}$$

Combining the above cases we deduce that

$$H_{pn}(\mathbf{g}) \geq (1 - \varepsilon - \frac{\lg(p^{-1}\varepsilon^{-2}\log n)}{\lg n})H(\mathbf{f}) - O(\varepsilon) - \frac{\varepsilon^{-2}\ln n}{pn} \ .$$

Setting $\varepsilon = p^{-1/2}n^{-1/6}$ we get

$$\begin{aligned} H_{pn}(\mathbf{g}) &\geq (1 - p^{-1/2}n^{-1/6} - \frac{\lg(n^{1/3}\log n)}{\lg n})H(\mathbf{f}) - O(p^{-1/2}n^{-1/6}) - O\left(\frac{\log n}{n^{2/3}}\right) \\ &\geq H(\mathbf{f})/2 - O(p^{-1/2}n^{-1/6}) \ . \end{aligned}$$

□

Therefore, by using an existing entropy estimation algorithm (e.g., [25]) to multiplicatively estimate $H(\mathbf{g})$ we have a constant factor approximation to $H(\mathbf{f})$ if $H(\mathbf{f}) = \omega(p^{-1/2}n^{-1/6})$. The next theorem follows directly from Proposition 1 and Lemma 10.

Theorem 5 *It is possible to approximate $H(\mathbf{f})$ up to a constant factor in $O(\text{polylog}(m,n))$ space if $H(\mathbf{f}) = \omega(p^{-1/2}n^{-1/6})$.*

6 Heavy Hitters

There are two common notions for finding heavy hitters in a stream: the F_1-heavy hitters, and the F_2-heavy hitters.

Definition 4 In the F_k-heavy hitters problem, $k \in \{1,2\}$ we are given a stream of updates to an underlying frequency vector f and parameters α, ε, and δ. The algorithm is required to output a set S of $O(1/\alpha)$ items such that: (1) every item i for which $f_i \geq \alpha(F_k)^{1/k}$ is included in S, and (2) any item i for which $f_i < (1-\varepsilon)\alpha(F_k)^{1/k}$ is not included in S. The algorithm is additionally required to output approximations f_i' with

$$\forall i \in S, \quad f_i' \in [(1-\varepsilon)f_i, (1+\varepsilon)f_i] \ .$$

The overall success probability should be at least $1-\delta$.

The intuition behind the algorithm for heavy hitters is as follows. Suppose an item i was an F_k heavy hitter in the original stream P, i.e. $f_i \geq \alpha(F_k)^{1/k}$. Then, by a Chernoff bound, it can be argued that with high probability, g_i the frequency of i in the sampled stream is close to pf_i. In such a case, it can be shown that i is also a heavy hitter in the sampled stream and will be detected by an algorithm that identifies heavy hitters on the sampled stream (with the right choice of parameters). Similarly, it can be argued that an item i such that $f_i < (1-\varepsilon)\alpha(F_k)^{1/k}$ cannot reach the required frequency threshold on the sampled stream, and will not be returned by the algorithm. We present the analysis below assuming that the heavy hitter algorithm on the sampled stream is the CountMin sketch. Other algorithms for heavy hitters can be used too, such as the Misra-Gries algorithm [33]; note that the Misra-Gries algorithm works on insert-only streams, while the CountMin sketch works on general update streams, with additions as well as deletions.

Theorem 6 *Suppose that*

$$F_1(P) \geq Cp^{-1}\alpha^{-1}\varepsilon^{-2}\log(n/\delta)$$

for a sufficiently large constant $C > 0$. There is a one pass streaming algorithm which observes the sampled stream L and computes the F_1 heavy hitters of the original stream P with probability at least $1-\delta$. This algorithm uses $O(\varepsilon^{-1}\log^2(n/(\alpha\delta)))$ bits of space.

Proof The algorithm runs the $\mathsf{CountMin}(\alpha', \varepsilon', \delta')$ algorithm of [15] for finding the F_1-heavy hitters problem on the sampled stream, for $\alpha' = (1-2\varepsilon/5)\cdot\alpha$, $\varepsilon' = \varepsilon/2$, and $\delta' = \delta/4$. We return the set S of items i found by CountMin, and we scale each of the f_i' by $1/p$.

Recall that g_i the frequency of item i in the sampled stream L. Then for sufficiently large $C > 0$ given in the theorem statement, for any i, by a Chernoff bound,

$$\Pr\left[g_i > \max\left\{p\left(1+\frac{\varepsilon}{5}\right)f_i,\ \frac{C}{2\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right\}\right] \leq \frac{\delta}{4n}.$$

By a union bound, with probability at least $1-\delta/4$, for all $i \in [n]$,

$$g_i \leq \max\left\{p\left(1+\frac{\varepsilon}{5}\right)f_i,\ \frac{C}{2\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right\}. \tag{7}$$

We also need the property that if $f_i \geq (1-\varepsilon)\alpha F_1(P)$, then $g_i \geq p(1-\varepsilon/5)f_i$. For such i, by the premise of the theorem we have

$$\mathbb{E}[g_i] \geq p(1-\varepsilon)\alpha F_1(P) \geq C(1-\varepsilon)\varepsilon^{-2}\log(n/\delta) .$$

Hence, for sufficiently large C, applying a Chernoff and a union bound is enough to conclude that with probability at least $1-\delta/4$, for all such i, $g_i \geq p(1-\varepsilon/5)f_i$.

We set the parameter δ' of CountMin to equal $\delta/4$, and so CountMin succeeds with probability at least $1-\delta/4$.

Also, $\mathbb{E}[[F_1(L)] = pF_1(P) \geq C\alpha^{-1}\varepsilon^{-2}(\log n/\delta)$, the inequality following from the premise of the theorem. By a Chernoff bound,

$$\Pr\left[\left(1-\frac{\varepsilon}{5}\right)pF_1(P) \leq F_1(L) \leq \left(1+\frac{\varepsilon}{5}\right)pF_1(P)\right] \geq 1-\frac{\delta}{4}.$$

By a union bound, all events discussed thus far jointly occur with probability at least $1-\delta$, and we condition on their joint occurrence in the remainder of the proof.

Lemma 11 *If $f_i \geq \alpha F_1(P)$, then*

$$g_i \geq (1-2\varepsilon/5)\cdot\alpha F_1(L) .$$

If $f_i < (1-\varepsilon)\alpha F_1(P)$, then

$$g_i \leq (1-\varepsilon/2)\alpha F_1(L) .$$

Proof Since $g_i \geq p(1-\varepsilon/5)f_i$ and also $F_1(L) \leq p(1+\varepsilon/5)F_1(P)$. Hence,

$$g_i \geq \frac{1-\varepsilon/5}{1+\varepsilon/5}\cdot\alpha F_1(L) \geq (1-2\varepsilon/5)\cdot\alpha F_1(L).$$

Next consider any i for which $f_i < (1-\varepsilon)\alpha F_1(P)$. Then

$$\begin{aligned}
g_i &\leq \max\left\{p\left(1+\frac{\varepsilon}{5}\right)(1-\varepsilon)\alpha F_1(P),\ \frac{C}{2\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right\}\\
&\leq \max\left\{\left(1-\frac{3\varepsilon}{5}\right)\alpha F_1(L),\ \frac{C}{2\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right\}\\
&\leq \max\left\{\left(1-\frac{\varepsilon}{2}\right)\alpha F_1(L),\ \frac{\alpha}{2}\cdot\mathbb{E}[F_1(L)]\right\}\\
&\leq \max\left\{\left(1-\frac{\varepsilon}{2}\right)\alpha F_1(L),\ \left(1+\frac{\varepsilon}{5}\right)\frac{\alpha}{2}F_1(L)\right\}\\
&\leq \left(1-\frac{\varepsilon}{2}\right)\alpha F_1(L).
\end{aligned}$$

□

It follows that by setting $\alpha' = (1-2\varepsilon/5)\cdot\alpha$ and $\varepsilon' = \varepsilon/2$, CountMin$(\alpha',\varepsilon',\delta')$ does not return any $i \in S$ for which $f_i < (1-\varepsilon)\alpha F_1(P)$, since for such i we have $g_i \leq (1-\varepsilon/2)\alpha F_1(L)$, and so $g_i < (1-\varepsilon/10)\alpha' F_1(L)$. On the other hand, for every $i \in S$ for which $f_i \geq \alpha F_1(P)$, we have $i \in S$, since for such i we have $g_i \geq \alpha' F_1(L)$.

It remains to show that for every $i \in S$, we have $f_i' \in [(1-\varepsilon)f_i, (1+\varepsilon)f_i]$. By the previous paragraph, for such i we have $f_i \geq (1-\varepsilon)\alpha F_1(P)$. By the above conditioning, this means

that $g_i \geq p(1-\varepsilon/5)f_i$. We will also have $g_i \leq p(1+\varepsilon/5)f_i$ if $p\left(1+\frac{\varepsilon}{5}\right)f_i \geq \frac{C}{2\varepsilon^2}\log\left(\frac{n}{\delta}\right)$. Since $f_i \geq (1-\varepsilon)\alpha F_1(P)$, this in turn holds if

$$F_1(P) \geq \frac{1}{2(1-\varepsilon)(1+\varepsilon/5)} \cdot Cp^{-1}\alpha^{-1}\varepsilon^{-2}\log\left(\frac{n}{\delta}\right) ,$$

which holds by the theorem premise provided ε is less than a sufficiently small constant. This completes the proof. □

Theorem 7 *Suppose that $F_2^{1/2} \geq Cp^{-3/2}\alpha^{-1}\varepsilon^{-2}\log(n/\delta)$ and $p = \tilde{\Omega}(m^{-1/2})$. There is a one pass streaming algorithm which observes the sampled stream L and computes $(\alpha, 1-p^{1/2}(1-\varepsilon))$ F_2-heavy hitters of the original stream with high probability.*

Proof The algorithm runs the $\mathsf{CountSketch}(\alpha', \varepsilon', \delta')$ algorithm [8] for finding the F_2-heavy hitters on the sampled stream, for appropriate α', ε', and δ' specified below. We return the set S of items i found by $\mathsf{CountSketch}$.

As before we can show that if $f_i \geq (1-\varepsilon)\alpha F_2^{1/2}$, then with probability at least $(1-\delta/4)$, $g_i \geq p(1-\varepsilon/5)f_i$. Next we bound the variance of $F_2(L)$. Since each g_i is drawn from a binomial distribution $\mathrm{Bin}(f_i, p)$ on f_i items with probability p,

$$\mathbb{E}\left[F_2(L)\right] = \sum_{i=1}^{n}\mathbb{E}\left[(g_i)^2\right] = \sum_{i=1}^{n}(p^2 f_i^2 + p(1-p)f_i) = p^2F_2(P) + p(1-p)F_1(P).$$

Moreover,

$$\mathbf{Var}[F_2(L)] = \sum_{i=1}^{n}\mathbf{Var}[(g_i)^2] \leq \sum_{i=1}^{n}\mathbb{E}\left[(g_i)^4\right] - (p^2f_i^2 + p(1-p)f_i)^2) \leq \sum_{i=1}^{n}\mathbb{E}\left[(g_i)^4\right] - p^4f_i^4.$$

It is known that the 4-th moment of $\mathrm{Bin}(f_i, p)$ is

$$f_ip(1-7p+7f_ip+12p^2-18f_ip^2+6f_i^2p^2-6p^3+11f_ip^3-6f_i^2p^3+f_i^3p^3),$$

and subtracting $p^4f_i^4$ from this, we obtain

$$f_ip-7f_ip^2+7f_i^2p^2+12f_ip^3-18f_i^2p^3+6f_i^3p^3-6f_ip^4+11f_i^2p^4-6f_i^3p^4+f_i^4p^4-f_i^4p^4,$$

which is $O(f_ip + f_i^2p^2 + f_i^3p^3)$. Hence, $\mathbf{Var}[F_2(L)] = O(pF_1 + p^2F_2(P) + p^3F_3(P))$. By Chebyshev's inequality,

$$\begin{aligned}\Pr\left[|F_2(L) - \mathbb{E}[F_2(L)]| \geq \varepsilon p^2F_2\right] &= O\left(\frac{pF_1+p^2F_2+p^3F_3}{\varepsilon^2p^4F_2^2}\right) = O\left(\frac{F_1}{\varepsilon^2pF_2^2} + \frac{1}{\varepsilon^2F_2} + \frac{pF_3}{\varepsilon^2F_2^2}\right)\\ &= O\left(\frac{1}{\varepsilon^2pF_2} + \frac{1}{\varepsilon^2F_2} + \frac{pF_2^{3/2}}{\varepsilon^2F_2^2}\right) = O\left(\frac{1}{\varepsilon^2pF_2} + \frac{1}{\varepsilon^2F_2} + \frac{p}{\varepsilon^2F_2^{1/2}}\right)\end{aligned}$$

Thus with probability at least $(1-\delta/4)$

$$\left(1-\frac{\varepsilon}{5}\right)pF_2^{1/2} \leq (F_2(L))^{1/2} \leq 2p^{1/2}F_2^{1/2}.$$

By union bound all events discussed so far jointly occur with probability at least $1-\delta$, and we condition on them occurring in the remainder of the analysis.

Suppose that $f_i \geq \alpha F_2^{1/2}$ in the original stream. Then

$$g_i \geq p(1-\varepsilon/5)f_i \geq \alpha F_2^{1/2} p(1-\varepsilon/5) \geq \alpha p^{1/2}(1-\varepsilon/5)F_2^{1/2}(L)$$

Next consider any i for which $f_i < (1-\varepsilon)p^{1/2}\alpha F_2^{1/2}$. Then

$$\begin{aligned}
g_i &\leq \max\left\{ p\left(1+\frac{\varepsilon}{5}\right)(1-\varepsilon)p^{1/2}\alpha F_2^{1/2}(P),\ \frac{C}{2\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right\} \\
&\leq \max\left\{ \left(1+\frac{\varepsilon}{5}\right)\left(1-\frac{4\varepsilon}{5}\right)\left(1-\frac{\varepsilon}{5}\right)p^{3/2}\alpha F_2^{1/2}(P),\ \frac{C}{2\varepsilon^2}\log\left(\frac{n}{\delta}\right)\right\} \\
&\leq \left(1-\frac{3\varepsilon}{5}\right)\left(1-\frac{\varepsilon}{5}\right)p^{3/2}F_2^{1/2}(P) \\
&\leq \left(1-\frac{\varepsilon}{2}\right)p^{1/2}\alpha(F_2(L))^{1/2}
\end{aligned}$$

It follows that by setting $\alpha' = (1-2\varepsilon/5)\cdot\alpha\cdot p^{1/2}$, $\delta' = \delta/4$, and $\varepsilon' = \varepsilon/10$, $\mathsf{CountSketch}(\alpha',\varepsilon',\delta')$ does not return any $i \in S$ for which $f_i < (1-\varepsilon)p^{1/2}\alpha F_2^{1/2}(P)$, since for such i we have $g_i \leq (1-\varepsilon/2)p^{1/2}\alpha(F_2(L))^{1/2}$. On the other hand, for every $i \in S$ for which $f_i \geq \alpha F_2^{1/2}$, we have $i \in S$, since for such i we have $g_i \geq \alpha'(F_2(L))^{1/2}$.

□

7 Conclusion

We presented small-space stream algorithms and lower bounds for estimating functions of interest when observing a random sample of the original stream. The are numerous directions for future work, and we mention some of them.

- As we have seen, our results imply time/space tradeoffs for several natural streaming problems. What other data stream problems have interesting time/space tradeoffs?
- Also, we have so far assumed that the sampling probability p is fixed, and that the algorithm has no control over it. Suppose this was not the case, and the algorithm can change the sampling probability in an adaptive manner, depending on the current state of the stream. Is it possible to get algorithms that can observe fewer elements overall and get the same accuracy as our algorithms? For which precise models and problems is adaptivity useful?
- It is also interesting to obtain matching space lower bounds for the case of estimating frequency moments.

References

1. Alon, N., Matias, Y., Szegedy, M.: The Space Complexity of Approximating the Frequency Moments. Journal of Computer and System Sciences **58**(1), 137–147 (1999)
2. Babcock, B., Datar, M., Motwani, R.: Sampling from a moving window over streaming data. In: Proc. ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 633–634 (2002)
3. Bar-Yossef, Z.: The complexity of massive dataset computations. Ph.D. thesis, University of California at Berkeley (2002)

4. Bar-Yossef, Z.: Sampling lower bounds via information theory. In: Proc. 35th Annual ACM Symposium on Theory Of Computing (STOC), pp. 335–344 (2003)
5. Barakat, C., Iannaccone, G., Diot, C.: Ranking flows from sampled traffic. In: Proc. ACM Conference on Emerging Network Experiment and Technology (CoNEXT), pp. 188–199 (2005)
6. Bhattacharyya, S., Madeira, A., Muthukrishnan, S., Ye, T.: How to scalably and accurately skip past streams. In: Proc. 23rd International Conference on Data Engineering (ICDE) Workshops, pp. 654–663 (2007)
7. Charikar, M., Chaudhuri, S., Motwani, R., Narasayya, V.R.: Towards estimation error guarantees for distinct values. In: Proc. 19th ACM Symposium on Principles of Database Systems (PODS), pp. 268–279 (2000)
8. Charikar, M., Chen, K., Farach-Colton, M.: Finding frequent items in data streams. Theoretical Computer Science **312**(1), 3–15 (2004)
9. Cisco Systems: Random Sampled NetFlow. `http://www.cisco.com/en/US/docs/ios/12_0s/feature/guide/nfstatsa.html`
10. Cohen, E., Cormode, G., Duffield, N.G.: Structure-aware sampling: Flexible and accurate summarization. Proceedings of the VLDB Endowment **4**(11), 819–830 (2011)
11. Cohen, E., Duffield, N.G., Kaplan, H., Lund, C., Thorup, M.: Efficient stream sampling for variance-optimal estimation of subset sums. SIAM J. Comput. **40**(5), 1402–1431 (2011)
12. Cohen, E., Duffield, N.G., Kaplan, H., Lund, C., Thorup, M.: Algorithms and estimators for summarization of unaggregated data streams. Journal of Computer and System Sciences **80**(7), 1214–1244 (2014)
13. Cohen, E., Grossaug, N., Kaplan, H.: Processing top-k queries from samples. Computer Networks **52**(14), 2605–2622 (2008)
14. Cormode, G., Garofalakis, M.: Sketching probabilistic data streams. In: Proc. 26th ACM International Conference on Management of Data (SIGMOD), pp. 281–292 (2007)
15. Cormode, G., Muthukrishnan, S.: An improved data stream summary: the count-min sketch and its applications. Journal of Algorithms **55**(1), 58–75 (2005)
16. Cormode, G., Muthukrishnan, S., Yi, K., Zhang, Q.: Optimal sampling from distributed streams. In: Proc. ACM Symposium on Principles of Database Systems (PODS), pp. 77–86 (2010)
17. Duffield, N.G., Lund, C., Thorup, M.: Properties and prediction of flow statistics from sampled packet streams. In: Proc. Internet Measurement Workshop, pp. 159–171 (2002)
18. Duffield, N.G., Lund, C., Thorup, M.: Estimating flow distributions from sampled flow statistics. IEEE/ACM Transactions on Networking **13**(5), 933–946 (2005)
19. Duffield, N.G., Lund, C., Thorup, M.: Priority sampling for estimation of arbitrary subset sums. Journal of the ACM **54**(6) (2007)
20. Efraimidis, P., Spirakis, P.G.: Weighted random sampling with a reservoir. Information Processing Letters **97**(5), 181–185 (2006)
21. Estan, C., Keys, K., Moore, D., Varghese, G.: Building a better netflow. In: Proc. ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM), pp. 245–256 (2004)
22. Estan, C., Varghese, G.: New directions in traffic measurement and accounting. In: Proc. ACM Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM), pp. 323–336 (2002)
23. Gibbons, P.B., Matias, Y.: New sampling-based summary statistics for improving approximate query answers. In: Proc. ACM SIGMOD International Conference on Management of Data, pp. 331–342 (1998)
24. Guha, S., Huang, Z.: Revisiting the direct sum theorem and space lower bounds in random order streams. In: Automata, Languages and Programming, 36th International Colloquium, ICALP (1), pp. 513–524 (2009)
25. Harvey, N.J.A., Nelson, J., Onak, K.: Sketching and streaming entropy via approximation theory. In: PRoc. 49th IEEE Conference on Foundations Of Computer Science (FOCS), pp. 489–498 (2008)
26. Hohn, N., Veitch, D.: Inverting sampled traffic. IEEE/ACM Transactions on Networking **14**(1), 68–80 (2006)
27. Indyk, P., Woodruff, D.P.: Optimal approximations of the frequency moments of data streams. In: Proc. 37th Annual ACM Symposium on Theory of Computing (STOC), pp. 202–208 (2005)
28. Jayram, T.S., McGregor, A., Muthukrishnan, S., Vee, E.: Estimating statistical aggregates on probabilistic data streams. ACM Transactions on Database Systems **33**, 26:1–26:30 (2008)
29. Kane, D.M., Nelson, J., Woodruff, D.P.: On the exact space complexity of sketching and streaming small norms. In: Proc. 21st ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1161–1178 (2010)
30. Lahiri, B., Tirthapura, S.: Stream sampling. In: L. Liu, M.T. Özsu (eds.) Encyclopedia of Database Systems, pp. 2838–2842. Springer US (2009)

31. McGregor, A. (ed.): Open Problems in Data Streams and Related Topics (2007). http://www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf
32. McGregor, A., Pavan, A., Tirthapura, S., Woodruff, D.: Space-efficient estimation of statistics over sub-sampled streams. In: Proc. 31st ACM Symposium on Principles of Database Systems (PODS), pp. 273–282 (2012)
33. Misra, J., Gries, D.: Finding repeated elements. Science of Computer Programming **2**(2), 143–152 (1982)
34. Rusu, F., Dobra, A.: Sketching sampled data streams. In: Proc. 25th IEEE International Conference on Data Engineering (ICDE), pp. 381–392 (2009)
35. Szegedy, M.: The dlt priority sampling is essentially optimal. In: Proc. Annual ACM Symposium on Theory of Computing (STOC), pp. 150–158 (2006)
36. Tirthapura, S., Woodruff, D.P.: Optimal random sampling from distributed streams revisited. In: Proc. International Symposium on Distributed Computing (DISC), pp. 283–297 (2011)
37. Vitter, J.S.: Random sampling with a reservoir. ACM Transactions on Mathematical Software **11**(1), 37–57 (1985)