Estimating small frequency moments of data stream: a characteristic function approach

Sumit Ganguly
IIT Kanpur
India

Purushottam Kar
IIT Kanpur
India

May 7, 2010

Abstract

We consider the problem of estimating the first moment of a data stream defined as $F_1 = \sum_{i \in \{1,2,\ldots,n\}} |f_i|$ to within $1 \pm \epsilon$-relative error with high probability. Several algorithms are well-known for this problem including the median estimator over p-stable sketches by Indyk [11], the geometric means estimator over p-stable sketches by Li [13] and the Hss sketch based algorithm in [8]. The current best algorithm is given by Kane, Nelson and Woodruff in [12] that uses space $O(\epsilon^{-2} \log(mM))$ and is proved to be space-optimal. In this paper, we present a novel, space-optimal algorithm for estimating F_p with an elementary analysis that is based on the characteristic function of stable distributions.

1 Introduction

There is an increasing need to compute over massive data streams that arrive rapidly and continuously. Sources of data range from network switches, sensor networks, web, transaction data etc.. Many applications over such data seek to monitor the health and integrity of the underlying system as reflected by the data (e.g., network, bridge, reactor, industrial plant), using very efficient algorithms that have guaranteed error-tolerance, invoking a "deep-analysis" algorithm only when the monitoring algorithms raise an alarm. A deep data analysis requires data storage and retrieval from secondary storage, which makes it an expensive operation. This has led to the design of *single-pass* online algorithms that process the arriving data in real time and are very efficient with respect to space or time (preferably both), typically requiring sub-linear (often only poly-logarithmic) space and time.

We can view a data stream computing model as follows. An input stream σ is abstracted as a potentially infinite sequence of records of the form (pos, i, v), where, $i \in \{1, 2, \ldots, n\} = [n]$ and $v \in \mathbb{Z}$ is the change to the frequency f_i of item i. The pos attribute is simply the sequence number of the record. Each input record (pos, i, v) changes $f_i \leftarrow f_i + v$. Thus, $f_i = \sum_{(\mathrm{pos},i,v)} v$, that is, f_i is the sum of the changes made to the frequency of i since the inception of the stream. The vector $f = [f_1, f_2, \ldots, f_n]^T$ is called the frequency vector of the stream. We assume that items come from the domain $[n] = \{1, 2, \ldots, n\}$, each stream update (pos, i, v) has $|v| \leq M$ and the size of the stream is m, that is, the number of records appearing in the stream.

The pth frequency moment is defined as $F_p = \sum_{i \in [n]} |f_i|^p$. The problem of estimating F_p, and in particular, the estimation of F_0, F_1 and F_2, have been fundamental to the development of data stream processing techniques and lower bounds [1, 11, 2, 15, 4, 7, 9, 3]. These problems have many applications. For example, in network monitoring F_1 sketches are used to detect differences in network traffic flows [6], F_1 and F_2 sketches are used for approximate histogram maintenance for database query optimization [10, 5] and F_p sketches can be used for fast approximations of document similarities over the web [13]. There are many basic primitives ranging from fast approximation of range-queries, finding heavy hitters, quantiles, etc. whose solutions are derived from F_p estimation techniques.

We will say that a randomized algorithm computes an ϵ-approximation to a real valued quantity L, provided, it returns $\hat{L}$ such that $|\hat{L} - L| < \epsilon L$, with probability that is at least some absolute constant strictly larger than $1/2$. In this paper, we consider the problem of estimating F_1 to within approximation factor of $1 \pm \epsilon$ and with probability at least some constant $c > 0.5$, where the probability is taken over the internal random bits used by the algorithm. Since prior work [1] shows that that any deterministic algorithm for 0.1-approximation of F_p, $p \geq 0$ requires $\Omega(n)$ space, we consider the problem of randomized ϵ-approximation of F_1.

1.1 Review: Previous work

[1] presents a seminal randomized sketch technique for ϵ-approximation of F_2 in the data streaming model using space $O(\epsilon^{-2} \log(mM))$ bits. Estimation of F_0 (i.e., number of $i \in [n]$ s.t. $|f_i| \neq 0$) was first considered by Flajolet and Martin in [7] and improved in [1, 9, 3]. Woodruff [16] presents an $\Omega(\epsilon^{-2})$ space lower bound for the problem of estimating F_p, for all $p \geq 0$. This is improved to $\Omega(\epsilon^{-2} \log(\epsilon^2 M))$ in [12]. Since the techniques for estimating F_p for $p > 2$ are substantially different from those used for estimating F_p for $0 < p < 2$, we do not review this line of work. The notation $X \sim D$ means that the random variable has probability distribution D. The term *i.i.d.* stands for independent and identically distributed family of random variables.

Indyk's estimator. Indyk in [11] presented the first algorithm for estimating F_1 using p-stable sketches, for $p \in (0, 2]$. A p-stable sketch is a linear combination $X = \sum_{i=1}^{n} a_i s_i$ where the s_i's are drawn independently from the p-stable distribution $\mathrm{St}(p, 1)$ with scale factor 1. By property of stable distributions, $X \sim S\big(p, (F_p(a))^{1/p}\big)$. For estimating F_1, keep $t = O(\frac{1}{\epsilon^2})$ independent Cauchy sketches (i.e., 1-stable) $X_1, X_2, \ldots, X_t$ and let $\hat{F}_1 = (4/\pi) \cdot \mathrm{median}_{r=1}^{t} |X_r|^q$. Then, $\hat{F}_1 \in (1 \pm \epsilon) F_1$ with probability $15/16$. Further, Indyk shows that for stable distributions it suffices to, (a) truncate the support of the distribution $\mathrm{St}(p, 1)$ beyond $(mM)^{O(1)}$, and, (b) consider the approximation to the continuous $\mathrm{St}(p, 1)$ distribution by discretizing it using into a grid with interval size $(mM/\epsilon)^{-O(1)}$.

To reduce the number of random bits required to maintain independent sketches, Nisan's pseudo-random generator (PRG) [14] is used for fooling space S bounded randomized machine computation. We can assume that the stream is ordered since the sketches are linear and therefore their values are independent of the order of item arrivals. For each element i, the stable random variables $s_i(u)$ for $u = 1, 2, \ldots, t$ are computed from the ith chunk of S random bits obtained from Nisan's generator that stretches a seed of length $S \log n$ to nS bits. The space used by Indyk's estimator is $S = O(\epsilon^{-2} \log(\epsilon^{-1} mM))$. The random seed size becomes $S \log n = O(\epsilon^{-2} \log(mM\epsilon^{-1}) \log(n))$ and this dominates the space requirement of the F_1 estimation algorithm. The time taken to obtain the ith random bit chunk is $O(\epsilon^{-2} \log(\epsilon^{-1}) (\log n))$ simple field operations on a field of size $(mM)^{O(1)}$. Kane, Nelson and Woodruff [12] observe that a seed length of $O(\log(\frac{mM}{\epsilon}) \log(n))$ suffices.

Li's estimator. Li [13] proposes the geometric means estimator for estimating F_p. Given p-stable sketches $X_u = \sum_{i\in[n]} f_i s_i(u)$, $u = 1, 2, \ldots, t$, the geometric means estimator is defined as

$$\hat{Y}_{p,t} = C(p, p/t)^{-t} \prod_{i=1}^{t} |X_i|^{p/t}.$$

where,

$$C(p,q) = \frac{2}{\pi}\Gamma\big(1-\frac{q}{p}\big)\Gamma(q)\sin\big(\frac{\pi q}{2}\big), \qquad -1 < q < p \ .$$

Li [13] proves that (i) the estimator is unbiased, that is, $\mathsf{E}\left[Y_{p,t}\right] = F_p$, and, (ii) $|\hat{Y}_{p,t} - F_p| < \epsilon F_p$ with prob. 1/8 provided, $t = \Omega(\epsilon^{-2})$.

Kane, Nelson, Woodruff's (KNW) estimator for F_p. Kane, Nelson and Woodruff [12] present two estimators for estimating F_p for $p \in (0, 2)$ that we denote by KNW-1 and KNW-2. Both these estimators use space that is tight with respect to the lower bounds, which was also improved in the same paper. The estimators view the computation of the p-stable sketches as the multiplication of the $t \times n$ random matrix A with the n-dimensional frequency vector f. Each $A_{i,j} \sim \mathcal{D}_p$, where, $\mathcal{D}_p$ is the discretized and truncated version of $\mathrm{St}(p, 1)$. However, unlike Indyk and Li's proposal to use fully independent $A_{i,j}$'s, the KNW-1 estimator requires just the following limited independence. (i) For each row value i, the column entries (i.e., $A_{i,j}$'s) are $O(\epsilon^{-p}\log^{3p}(1/\epsilon))$-wise independent, and, (ii) the rows of A are pair-wise independent. This can be achieved using a random seed of size $O(t \log(mM)) = O\big(\epsilon^{-p}\log^{3p}(1/\epsilon)\log(mM)\big)$. The update processing time requirement is $O(\epsilon^{-2-p}\log^{3p}(1/\epsilon))$. The KNW-2 estimator further reduces the independence requirement among the variates in a single row of A to $\log(\epsilon^{-1})/\log\log(\epsilon^{-1})$. This reduces the estimation time to $O\big(\epsilon^{-2}(\log\epsilon^{-1})^2/(\log\log\epsilon^{-1})\big)$ simple operations on fields of size $(mM)^{O(1)}$.

Hss estimator. An estimator for F_p based on the HSS technique was presented in [8] for estimating F_p. Though it uses sub-optimal space $O(\epsilon^{-2-p}(\log(mM)^2(\log n))$, it has the best update processing time so far, namely, $O(\log^2(mM))$.

Contributions We present a novel, space-optimal algorithm for estimating F_p that uses space $O(\epsilon^{-2}\log(mM))$. The algorithm has a simpler and elementary analysis as compared to the space-optimal estimators of [12] and directly utilizes the characteristic function of stable distributions.

2 A space-optimal estimator for F_p

Let $s = 256 \cdot e^5 \cdot \epsilon^{-2}$. Keep p-stable sketches $X_1, X_2, \ldots, X_s$, where, $X_i = \sum_{j=1}^{n} f_j s_{i,j}$, and $s_{i,j} \sim \mathrm{St}(p, 1)$. The family $\{s_{i,j}\}_{j=1}^{n}$, for each fixed i, is assumed to be $\log(1/\epsilon)$-wise independent. The random seeds generating the families $\{s_{i,j}\}$ and $\{s_{i',j}\}$, for $i \neq i'$, need only be pair-wise independent. Define

$$C_s(t) = \frac{1}{s}(\cos(tX_1) + \ldots \cos(tX_s)) \quad \text{and} \quad \hat{F}_p(t) = \frac{1}{t^p}\log\frac{1}{C_s(t)} \ . \tag{1}$$

The estimation procedure is as follows. The estimate 0 is returned if all the X_i's are 0. Otherwise, values of $t = 1, 2^{-1/p}, 2^{-2/p}, \ldots, 2^{(-3+\lceil\log(mM)\rceil)/p}$ are chosen until one is found that satisfies the following condition.

$$\big(1+\frac{\epsilon}{16}\big)e^{-1} \le C_s(t) \le \big(1-\frac{\epsilon}{16}\big)e^{-1/8} \quad . \tag{2}$$

If no such t is found to satisfy (2) then 0 is returned. Otherwise, the estimate $\hat{F}_p(t)$ given by (4) is returned.

Analysis. We first prove the correctness of the estimator assuming full-independence of the $s_{i,j}$'s for each fixed i. We will then relax the independence required.

Lemma 2.1 *Suppose the family of random variables $\{s_{i,j}\}_{j=1}^n$ is fully independent and for each $i \neq i'$, the random seeds generating the families $\{s_{i,j}\}$ and $\{s_{i',j}\}$, for $i \neq i'$, are pair-wise independent. Let $s = 256e^5\epsilon^{-2}$, $\epsilon \le 1$ and $F_p \ge 1$.*

1. *If t satisfies (2) then $\hat{F}_p(t) \in (1 \pm 8\epsilon/15)F_p$ with probability at least $1 - 2e^{-3} - e^{-4.75} \ge 0.89$.*
2. *If $F_p \ge 1$, then a value of t satisfying (2) can be found with probability at least $1 - 10/s \ge 0.9997$.*
3. *If $X_j = 0$ for $j = 1, 2, \ldots, s$, then, $F_p = 0$ with probability 1.*

Proof Let X be a p-stable sketch, $X = \sum_{j=1}^n f_j s_j$, where, $s_j \sim \mathrm{St}(p,1)$ and *i.i.d.*. The characteristic function of X satisfies

$$\mathsf{E}\left[e^{itX}\right] = \mathsf{E}\left[\cos(tX)\right] + i\mathsf{E}\left[\sin(tX)\right] = e^{-F_p|t|^p} \quad . \tag{3}$$

Thus we have $\mathsf{E}\left[\cos(tX)\right] = e^{-F_p|t|^p}$. Also, clearly $\mathsf{Var}\big[\cos(tX)\big] \le 1$.
Since $C_s(t)$ is the average of s pair-wise independent observations $\cos(tX)$,

$$\mathsf{E}\left[C_s(t)\right] = e^{-F_p|t|^p} \text{ and } \mathsf{Var}\big[C_s(t)\big] \le 1/s \quad . \tag{4}$$

We now make the following three claims.

Claim 1. If $|t|^p F_p \le 1$ then, $\mathsf{Pr}\left[|C_s(t) - e^{-|t|^p F_p}| \ge \frac{\epsilon}{16} e^{-F_p|t|^p}\right] \le e^{-3}$.

Proof of Claim 1. Let $|t|^p F_p \le 1$.

$$\begin{aligned}\mathsf{Pr}\left[|C_s(t) - e^{-|t|^p F_p}| \ge \frac{\epsilon}{16} e^{-F_p|t|^p}\right] &= \mathsf{Pr}\left[|C_s(t) - \mathsf{E}\left[C_s(t)\right]| \ge \frac{\epsilon}{16} e^{-|t|^p F_p}\right] \\ &\le \frac{\mathsf{Var}\big[C_s(t)\big]}{(\epsilon/16)^2 e^{-2F_p|t|^p}} \le \frac{1}{s(\epsilon/16)^2 e^{-2}} \le \frac{1}{e^3} \quad .\end{aligned}$$

The first inequality follows since $\mathsf{E}\left[C_s(t)\right] = e^{-|t|^p F_p}$, the second inequality follows from Chebychev's inequality and the third inequality follows by (4) and since $|t|^p F_p \le 1$, and the final inequality follows since $s = 256\epsilon^{-2}e^5$.

Claim 2. If $|t|^p F_p < 1/8$ then $\mathsf{Pr}\left[C_s(t) < \big(1 - \frac{\epsilon}{16}\big)e^{-1/8}\right] \le e^{-4.75}$.

Proof of Claim 2. Let $|t|^p F_p < 1/8$. Then,

$$\begin{aligned}&\mathsf{Pr}\left[C_s(t) < \big(1-\frac{\epsilon}{16}\big)e^{-1/8}\right] = \mathsf{Pr}\left[e^{-F_p|t|^p} - C_s(t) > e^{-F_p|t|^p} - (1 - \frac{\epsilon}{16})e^{-1/8}\right] \\ &\le \mathsf{Pr}\left[\big|C_s(t) - \mathsf{E}\left[C_s(t)\right]\big| > \frac{\epsilon}{16}e^{-1/8}\right] \le \frac{1}{s(\epsilon/16)^2 e^{-1/4}} = e^{-4.75} \quad .\end{aligned}$$

The first equality follows from $\mathsf{E}[C_s(t)] = e^{-|t|^p F_p}$, the second inequality follows since $|t|^p F_p < 1/8$ and so $e^{-|t|^p F_p} > e^{-1/8}$, the third inequality follows from Chebychev's inequality and since $\mathsf{Var}\big[C_s(t)\big] \le 1/s$, and the last inequality follows by substituting $s = 256\epsilon^{-2}e^5$.

Claim 3. If $|t|^p F_p > 1$ then $\mathsf{Pr}\left[C_s(t) > \big(1+\frac{\epsilon}{16}\big)e^{-1}\right] \le e^{-3}$.

Proof of Claim 3. Let $|t|^p F_p > 1$. Then,

$$\begin{aligned}\mathsf{Pr}\left[C_s(t) > \big(1+\frac{\epsilon}{16}\big)e^{-1}\right] &= \mathsf{Pr}\left[C_s(t) - \mathsf{E}[C_s(t)] > \big(1+\frac{\epsilon}{16}\big)e^{-1} - e^{-|t|^p F_p}\right] \\ &\le \mathsf{Pr}\left[|C_s(t) - \mathsf{E}[C_s(t)]| > \frac{\epsilon}{16}e^{-1}\right] \le \frac{1}{se^{-2}(\epsilon/16)^2} = e^{-3} \ .\end{aligned}$$

The arguments are the same as in Claim 2, except that the second inequality is obtained using the fact that $|t|^p F_p > 1$ implies $e^{-|t|^p F_p} < e^{-1}$.

We now return to the proof of the lemma. Define the event $\textsc{Good}(t)$ to be

$$\textsc{Good}(t) \equiv \frac{1}{8} \le |t|^p F_p \le 1 \text{ and } C_s(t) \in \big(1 \pm \frac{\epsilon}{16}\big)e^{-|t|^p F_p} \ .$$

By Claims 1 through 3, using union bound,

$$\mathsf{Pr}[\textsc{Good}(t)] \ge 1 - 2e^{-3} - e^{-4.75} \ge 0.89 \ .$$

We will now assume that $\textsc{Good}(t)$ holds. Expanding $\log(1/C_s(t))$ around $e^{-|t|^p F_p}$ using Taylor's series, there exists $\lambda \in [e^{-|t|^p F_p}, C_s(t)]$ such that

$$\left|\log\frac{1}{C_s(t)} - |t|^p F_p\right| = \left|\frac{C_s(t) - e^{-|t|^p F_p}}{\lambda}\right| \le \frac{\epsilon}{16}\frac{e^{-|t|^p F_p}}{e^{-|t|^p F_p}(1-\epsilon/16)} = \frac{\epsilon}{16(1-\epsilon/16)} \tag{5}$$

where, the first inequality is obtained by Taylor's series and the second inequality follows since $\textsc{Good}(t)$ implies that $|C_s(t) - e^{-|t|^p F_p}| \le (\epsilon/16)e^{-|t|^p F_p}$. Therefore,

$$\big|\hat{F}_p(t) - F_p\big| = \left|\frac{1}{|t|^p}\log\frac{1}{C_s(t)} - F_p\right| \le \frac{\epsilon}{|t|^p(1-\epsilon/16)} \le \frac{8\epsilon F_p}{16(1-\epsilon/16)} \le \frac{8\epsilon F_p}{15} \ .$$

Here the first equality follows from the definition of $\hat{F}_p(t)$ from (4), the second inequality follows (5), the third inequality follows since $\textsc{Good}(t)$ implies $|t|_p^F \ge 1/8$ and the final inequality follows since $\epsilon \le 1$. Thus

$$\begin{aligned}\mathsf{Pr}\left[\big|\hat{F}_p(t) - F_p(t)\big| < 8\epsilon F_p/15\right] &= \mathsf{Pr}\left[\big|\hat{F}_p(t) - F_p(t)\big| < 8\epsilon F_p/15 \mid \textsc{Good}(t)\right]\mathsf{Pr}[\textsc{Good}(t)] \\ &= 1 \cdot \mathsf{Pr}[\textsc{Good}(t)] \ge 0.89 \ .\end{aligned}$$

This proves the first statement of the lemma. We now prove the second statement of the lemma.

Claim 4. Suppose that $\frac{1}{8F_p} \le |t|^p \le \frac{1}{F_p}$. Then,

$$\mathsf{Pr}\left[\big(1+\frac{\epsilon}{16}\big)e^{-1} \le C_s(t) \le \big(1-\frac{\epsilon}{16}\big)e^{-1/8}\right] \ge 1 - \frac{10}{s} \ .$$

Proof of Claim 4. Suppose $|t|^p F_p \le 1$. Then,

$$\begin{aligned}\mathsf{Pr}\left[C_s(t) > \big(1-\frac{\epsilon}{16}\big)e^{-1/8}\right] &= \mathsf{Pr}\left[C_s(t) - \mathsf{E}[C_s(t)] > \big(1-\frac{\epsilon}{16}\big)e^{-1/8} - e^{-1}\right] \\ &\le \frac{1}{s\big((1-\frac{\epsilon}{16})e^{-1/8} - e^{-1}\big)^2} \le \frac{5}{s} \ .\end{aligned}$$

Suppose $|t|^p F_p \geq \frac{1}{8}$. Then,

$$\mathsf{Pr}\left[C_s(t) < \left(1+\frac{1}{16}\right)\epsilon^{-1}\right] = \mathsf{Pr}\left[\mathsf{E}\left[C_s(t)\right] - C_s(t) \geq e^{-1/8} - \left(1+\frac{1}{16}\right)e^{-1}\right]$$
$$\leq \frac{1}{s\left(e^{-1/8} - \left(1+\frac{1}{16}\right)e^{-1}\right)^2} \leq \frac{5}{s} \ .$$

The arguments used are similar to previous claims. By union bound, Claim 4 is proved.

Finally, if $F_p = 0$ then $X_j = 0$ for each j. Conversely, if $F_p \neq 0$, then, the probability that $X_1, \ldots, X_s$ are all zeros has measure 0. ∎

We can now show that $O(\log(\epsilon^{-1})/\log\log(\epsilon^{-1}))$-wise independence of the stable variables suffices. Let $\mathsf{E}\left[\cos(tX)\right]$ and $\mathsf{E}_{2a}\left[\cos(tX)\right]$ denote the expectations when the stable variables s_j's forming X are respectively fully independent and $2a$-wise independent.

Lemma 2.2 *For $\epsilon < 1/2$, $|\mathsf{E}\left[\cos(tX)\right] - \mathsf{E}_{2a}\left[\cos(tX)\right]| \leq \epsilon^4$.*

Proof For any tX,

$$\left|\cos(tX) - \sum_{r=0}^{2a-2} \frac{(-1)^r (tX \mod 2\pi)^{2r}}{(2r)!}\right| \leq \frac{(2\pi)^{2a}}{(2a)!} \leq \frac{\epsilon^4}{2}, \quad \text{if } 2a = \frac{8\pi e \log \epsilon^{-1})}{\log\log(\epsilon^{-1})}.$$

Since $X = \sum_j f_j s_j$ and the s_j's are $2a$-wise independent, it follows that $\mathsf{E}\left[(tX)^r \mod 2\pi\right] = \mathsf{E}_{2a}\left[(tX)^r \mod 2\pi\right]$, for $r \leq 2a$, since X^r only involves product terms with at most r distinct s_j's. Therefore, $|\mathsf{E}\left[\cos(tX)\right] - \mathsf{E}_{2a}\left[\cos(tX)\right]| \leq 2 \cdot \frac{\epsilon^4}{2} = \epsilon^4$. ∎

By the above analysis, the number of random bits used by the estimator is $O(\frac{\log(\epsilon^{-1}}{\log\log(\epsilon^{-1})}\log(mM))$. The space requirement of the estimator is therefore $O(\epsilon^{-2}\log(mM))$ which matches the lower bound in [12]. The update time requirement is $O\left(\epsilon^{-2}\frac{\log(\epsilon^{-1})}{\log\log(\epsilon^{-1})}\right)$ matching that of the KNW-2 estimator [12]. In the next section, we present an estimator for F_1 with a significantly improved time requirement.

3 Conclusion

We present a novel space-optimal algorithm for estimating F_p over data streams to within multiplicative error factor of $1 \pm \epsilon$ for $p \in (0, 2]$. The algorithm has an elementary analysis as compared to previous space-optimal algorithms and is based on the characteristic function of stable distributions.

References

[1] Noga Alon, Yossi Matias, and Mario Szegedy. "The space complexity of approximating frequency moments". *J. Comp. Sys. and Sc.*, 58(1):137–147, 1998. Preliminary version appeared in Proceedings of *ACM STOC* 1996, pp. 1-10.

[2] Z. Bar-Yossef, T.S. Jayram, R. Kumar, and D. Sivakumar. "An information statistics approach to data stream and communication complexity". In *Proceedings of ACM STOC*, pages 209–218, Princeton, NJ, 2002.

[3] Z. Bar-Yossef, T.S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. "Counting distinct elements in a data stream". In *Proceedings of International Workshop on Randomization and Computation (RANDOM)*, Cambridge, MA, 2002.

[4] A. Chakrabarti, S. Khot, and X. Sun. "Near-Optimal Lower Bounds on the Multi-Party Communication Complexity of Set Disjointness". In *Proceedings of International Conference on Computational Complexity (CCC)*, Aarhus, Denmark, 2003.

[5] G. Cormode, P. Indyk, N. Koudas, and S. Muthukrishnan. "Fast mining of massive tabular data via approximate distance computations". In *Proceedings of IEEE International Conference on Data Engineering*, pages 605–, 2002.

[6] Joanne Feigenbaum, Sampath Kannan, Martin Strauss, and M. Viswanathan. "An Approximate L^1-Difference Algorithm for Massive Data Streams". In *Proceedings of IEEE FOCS*, pages 501–511, New York, NY, October 1999.

[7] P. Flajolet and G.N. Martin. "Probabilistic Counting Algorithms for Database Applications". *J. Comp. Sys. and Sc.*, 31(2):182–209, 1985.

[8] Sumit Ganguly and Graham Cormode. "On Estimating Frequency Moments of Data Streams". In *Proceedings of International Workshop on Randomization and Computation (RANDOM)*, 2007.

[9] P. B. Gibbons and S. Tirthapura. "Estimating simple functions on the union of data streams". In *Proceedings of ACM SPAA*, pages 281–291, Heraklion, Crete, Greece, 2001.

[10] Anna Gilbert, Sudipto Guha, Piotr Indyk, Y. Kotidis, S. Muthukrishnan, and Martin Strauss. "Fast Small-space Algorithms for Approximate Histogram Maintenance". In *Proceedings of ACM STOC*, pages 152–161, 2002.

[11] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006. Preliminary Version appeared in Proceedings of IEEE FOCS 2000, pages 189-197.

[12] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. "On the Exact Space Complexity of Sketching and Streaming Small Norms". In *Proceedings of ACM Symposium on Discrete Algorithms (SODA)*, 2010.

[13] Ping Li. Estimators and tail bounds for dimension reduction in ℓ_α $(0 < \alpha \leq 2)$ using stable random projections. In *Proceedings of ACM Symposium on Discrete Algorithms (SODA)*, pages 10–19, 2008.

[14] N. Nisan. "Pseudo-Random Generators for Space Bounded Computation". In *Proceedings of ACM Symposium on Theory of Computing STOC*, pages 204–212, May 1990.

[15] M. Saks and X. Sun. "Space lower bounds for distance approximation in the data stream model". In *Proceedings of ACM STOC*, 2002.

[16] David P. Woodruff. "Optimal space lower bounds for all frequency moments". In *Proceedings of ACM Symposium on Discrete Algorithms (SODA)*, pages 167–175, 2004.