The Data Stream Space Complexity of Cascaded Norms

T.S. Jayram[†] and David P. Woodruff[†]
[†] *IBM Almaden Research Center*

Abstract— We consider the problem of estimating *cascaded aggregates* over a matrix presented as a sequence of updates in a data stream. A cascaded aggregate $P \circ Q$ is defined by evaluating aggregate Q repeatedly over each row of the matrix, and then evaluating aggregate P over the resulting vector of values. This problem was introduced by Cormode and Muthukrishnan, PODS, 2005 [CM].

We analyze the space complexity of estimating *cascaded norms* on an $n \times d$ matrix to within a small relative error. Let L_p denote the p-th norm, where p is a non-negative integer. We abbreviate the cascaded norm $L_k \circ L_p$ by $L_{k,p}$.

(1) For any constant $k \geq p \geq 2$, we obtain a 1-pass $\widetilde{O}(n^{1-2/k}d^{1-2/p})$-space algorithm for estimating $L_{k,p}$. This is optimal up to polylogarithmic factors and resolves an open question of [CM] regarding the space complexity of $L_{4,2}$. We also obtain 1-pass space-optimal algorithms for estimating $L_{\infty,k}$ and $L_{k,\infty}$.
(2) We prove a space lower bound of $\Omega(n^{1-1/k})$ on estimating $L_{k,0}$ and $L_{k,1}$, resolving an open question due to Indyk, IITK Data Streams Workshop (Problem 8), 2006.

We also resolve two more questions of [CM] concerning $L_{k,2}$ estimation and block heavy hitter problems. Ganguly, Bansal and Dube (FAW, 2008) claimed an $\widetilde{O}(1)$-space algorithm for estimating $L_{k,p}$ for any $k,p \in [0,2]$. Our lower bounds show this claim is incorrect.

1. Introduction

The recent explosion in the processing of terabyte-sized data sets has led to significant scientific advances as well as competitive advantages for economic entities. With the widespread adoption of information technology in healthcare, and in the tracking of individual clicks over the internet, massive data sets have become increasingly important on a societal and personal level. The constraints imposed by processing this massive data have inspired highly successful new paradigms, such as the data stream model, in which a processor makes a quick "sketch" of its input data in a single pass and is able to extract important statistical properties of the data. This has yielded efficient algorithms for several classical problems in the area including frequency-based statistics, ranking-based statistics, metric norms, and similarity measures (see [24] for a survey), and a complementary rich set of lower-bound techniques and results [2], [6], [27].

Two classical problems in the foundations of processing massive data sets are *frequency moments* and *norms*. For a stream X with general updates, let $w_a(X)$ denote the total weight of an item a induced by the weighted increments and decrements to a. Let the *k-th frequency moment* $F_k(X) \triangleq \sum_a |w_a(X)|^k$ and the *k-th norm* $L_k(X) \triangleq F_k^{1/k}$. For simplicity, we describe the results using norms; the analogous statements for frequency moments can be obtained via suitable manipulations of exponents. Special cases include distinct elements (L_0), Euclidean norms (L_2), the mode (L_∞) and the closely related "heavy-hitters" problem, that have all been thoroughly studied [15], [2], [7], [6], [11], [19]. Estimating L_k for $k > 2$ has applications in statistics to the *skewness* and *kurtosis* of a random variable, that provide a measure of asymmetry of a distribution (see http://en.wikipedia.org/wiki/Kurtosis_risk).

Although norms are useful measure for *single-attribute* aggregation, most applications deal with multi-dimensional data. Here, the real insights are obtained by slicing the data multiple times and applying *several aggregate measures in a cascaded fashion*, e.g., volatility in the stock market, IP traffic [13], r-means and r-median problems in computational geometry [14], approximating various matrix norms, and product norms in metric spaces [4]. A common query, involving two levels of aggregation, was introduced by Cormode and Muthukrishnan [13]:

Definition 1 (Cascaded Aggregates)**.** Consider a stream X consisting of integer updates to items in a matrix $[n] \times [d]$. We assume that the maximum update in magnitude, the stream length, n and d are all polynomially related. Let W denote the matrix whose (i,j)-th entry is $w_{ij}(X)$. Given two aggregate operators P and Q, the cascaded aggregate $P \circ Q$ is obtained by first applying Q to each row of A, and then applying P to the resulting vector of values. Abusing notation, we also apply $P \circ Q$ to X and denote $(P \circ Q)(X) = P(Q(X_1), Q(X_2), \ldots, Q(X_n))$, where X_i, for each $i \in [n]$, denotes the sub-stream of X corresponding to updates to item (i,j) for all $j \in [d]$. We abbreviate the cascaded norm $L_k \circ L_p$ by $L_{k,p}$.

The focus in [13] was mostly on the case $P \circ L_0$ for different choices of P. For estimating $L_{2,0}$, under the restriction that the updates are nonnegative, they

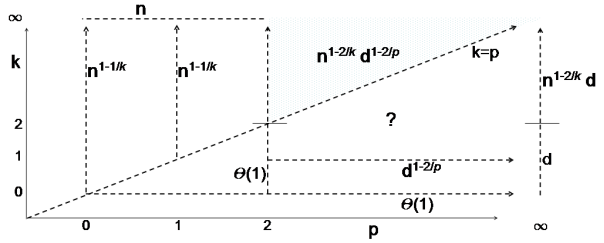

Figure 1. The optimal space complexity (upto $\widetilde{O}(1)$ factors) of estimating $L_{k,p}$ of an $n \times d$ matrix where k and p are integers. The upper bound for $k \geq p$, where $p = 0$ is due to [13] and $p = 1$ to [23], and the bound for $k = 1$ and $p > 2$ can be obtained using the techniques of [3].

gave an algorithm using $\widetilde{O}(\sqrt{n})$ space[1]. They left open the problem of whether the space bound could be improved. Under the same restriction of non-negative updates, Ganguly *et al.* [16] obtained an $\widetilde{O}(1)$ space algorithm for estimating $L_{k,p}$ where $0 \leq k \leq 1$ and $0 \leq p \leq 2$. One can estimate $L_{1,p}$ for $p > 2$ using space $\widetilde{O}(d^{1-2/p})$ by applying Theorem 1.5 of [3][2]. Also, Monemizadeh and the second author [23] gave an algorithm for estimating $L_{k,1}$ for $k \geq 1$ using space $\widetilde{O}(n^{1-1/k})$. Both of these algorithms can handle positive as well as negative updates.

In this paper, we study the space complexity of estimating cascaded norms $L_{k,p}$ to within a $(1 \pm \varepsilon)$-factor, where k and p are assumed to be non-negative integers. We obtain a near-complete characterization of the problem for a large regime of parameter values. Our results are summarized in Figure 1.

The main result in the paper, and also technically the most involved, is the following: for any constants $k \geq p \geq 2$, we obtain a 1-pass $\widetilde{O}(n^{1-2/k}d^{1-2/p})$-space algorithm for estimating $L_{k,p}$. In particular, we show that the space complexity of estimating $L_{4,2} = F_2 \circ F_2$ is $\widetilde{\Theta}(\sqrt{n})$, thus resolving an open question of Cormode and Muthukrishnan. The space complexity of our algorithm is optimal up to $\widetilde{O}(1)$ factors.

Since $L_{k,p}^k = F_{k/p} \circ F_p$, overloading notation, we will consider the problem of estimating $F_{k,p} \triangleq F_k \circ F_p$, $k \geq 1$, $p \geq 2$, using space $\widetilde{O}(n^{1-2/(kp)}d^{1-2/k})$. Naively estimating F_p on every row i using space-optimal F_p estimation algorithms uses too much space. To overcome this hurdle, we design two generic 1-pass procedures that have potential applications to other F_p-type computations.

[1]The tilde notation shall hide poly$(\log(n)/\varepsilon)$ factors.

[2]One needs to slightly adjust their data structure to directly obtain the parameter "γ" mentioned there, as well as use the L_p-sketch of [19] in the buckets of their data structure.

Our first procedure is a space efficient *F_p-sampling algorithm*, $p \geq 2$, for obtaining a large number of samples (with replacement) according to an approximate histogram for F_p. While the F_p estimation algorithm of [19] also yields an approximate frequency histogram, the variance of the estimator is too large, and the samples obtained from the approximate histogram are not sufficient. Further, repeating the procedure will result in a huge blowup in space. We add new ingredients to limit the space used to generate the samples. In particular, we resort to another subsampling procedure to handle levels in the histogram that have many more items than the expected number of samples needed from this level. For $F_{k,p}$ estimation, the F_p-sampling algorithm can be used to return row ids according to their approximate F_p value by ignoring the matrix structure. If we also had the F_p values for the chosen row ids, we could have fed them into an appropriate F_k estimator. In effect, this estimator is the one used in the F_k algorithm in [2]. However, the sampling guarantees are not sufficient to estimate the F_p value of the chosen row ids, thereby incurring an additional pass.

To avoid the second pass, we design a new *oracle F_k-estimation algorithm*, $k \geq 1$, that is based only on obtaining sample items according to their weight but whose weights themselves are not available. Technically, we need the following stronger oracle: for a parameter j, the oracle forms a sub-stream by including each item in the domain independently with probability 2^{-j}, and returns a large number of samples according to their weight in the substream. By additionally estimating the L_1 norm of the substream (which via known algorithms uses small space), we show how to obtain an appropriate F_k estimator. The analysis is quite complicated because the sampling is not exact and items may be misclassified if their weights are too close to the boundaries of the different weight levels.

A few more comments regarding our $L_{k,p}$ contribution are worth mentioning. In the literature, there is another algorithm for L_k estimation in a data stream due to Bhuvanagiri *et al.* [8], which achieves a much smaller space by a poly$((\log n)/\varepsilon)$ factor. However, it seems this latter algorithm cannot be easily modified to obtain an approximate histogram. We are not aware of other work which builds on the histogram of [19].

Next, we consider the problem of estimating $L_{k,p}$ where $k > 0$ and $p \in \{0,1\}$. As mentioned earlier, [13] gave an algorithm for a restricted case of $L_{2,0}$-estimation using $\widetilde{O}(\sqrt{n})$ space. Indyk [1, Problem 8] asked whether any non-trivial lower bounds can be

shown for this problem. We prove a space lower bound of $\Omega(n^{1-1/k})$ for this problem, which matches the upper bound of [13] with $k = 2$ and $p = 0$.

Our proof involves a new technique for information complexity [9], [6] for an appropriate communication problem. Via the information complexity direct sum theorem, we reduce it to the following problem: given two input vectors in the Hamming cube $\{0,1\}^d$ of dimension d, show that any communication protocol that can distinguish close instances (distance at most 1) from far instances (distance d) must have information complexity $\Omega(1/d)$. The caveat, as in previous applications of this paradigm, is that the support of the distribution is only on the close instances, so just the correctness of the protocol does not immediately yield the desired lower bound. Bar-Yossef *et al.* [6] manage to prove such a lower bound for L_∞ by resorting to the Hellinger distance in statistics. Using the Pythagorean property of communication protocols, the information cost is bounded in terms of distances between pairs of close and far instances, as needed. Adapting this proof for our case yields a weaker lower bound of $\Omega(1/d^2)$, which is not useful for $L_{k,0}$. We show how to exploit the Euclidean nature of Hellinger distance by employing the "short diagonals" property of points in Euclidean space. This property has been used previously to give non-embeddability results for metric spaces but this is the first application, to our knowledge, to communication complexity and data streams. Following our work, further applications of the "information geometry" of communication protocols have been shown to get improved lower bounds for edit distance [5], and to simplify the proof for the multiparty number-in-hand information complexity of AND [20].

Next, for $L_{k,2}$ for any $k > 0$, we obtain a 1-pass space-optimal algorithm for estimating $L_{k,2}$. Our techniques also allow us to find all rows whose L_2 norm is at least a constant $\phi > 0$ fraction of $L_{1,2}$ in $\widetilde{O}(1)$ space, i.e., to solve the "heavy hitters" problem for rows of the matrix weighted by L_2 norm. These results resolve two more open questions of Cormode and Muthukrishnan. Finally, for $k \geq 1$, we obtain 1-pass space-optimal algorithms for $L_{\infty,k}$ and $L_{k,\infty}$.

We note that previously, Ganguly, Bansal, and Dube [16] claimed an $\widetilde{O}(1)$-space algorithm for estimating $L_{k,p}$ for any k, p in $[0,2]$. Our lower bounds, as well as a reduction from multiparty set disjointness, show that this claim is incorrect.

Reducing Randomness: We describe our algorithms using random oracles, i.e., they have access to unlimited randomness including the use of continuous distributions. This assumption can be eliminated by the use of pseudorandom generators (PRGs) [25], similar to Indyk [18]. We give the standard transformation in the full version of this paper.

Proposition 2 (Hölder's inequality)**.** *Given a stream X of updates to D distinct items, (a) $F_2(X) \leq D^{1-2/p} \cdot F_p(X)^{2/p}$, $p \geq 2$; (b) $F_1(X) \leq D^{1-1/k} \cdot F_k(X)^{1/k}$, $k \geq 1$.*

2. Cascaded Frequency Moments

In this section, we show that the space complexity of estimating $L_{k,p}$, when $k \geq p \geq 2$, is an optimal $\widetilde{O}(n^{1-2/k}d^{1-2/p})$. Due to lack of space, many proofs are omitted from this version.

The lower bound follows via a simple reduction from multiparty set disjointness. Specifically, the inputs are $t = 2n^{1/k}d^{1/p}$ subsets of $[n] \times [d]$ such that on a NO instance, the sets are pairwise disjoint, and on a YES instance there exists (i, j) such that the intersection of every distinct pair of sets equals (i, j). The sets can be concatenated to form a stream input X for $L_{k,p}$ in a standard manner. For a NO instance, $w_{ij} \in \{0,1\}$ for every i, j. Therefore $L_{k,p}(X) \leq (\sum_i d^{k/p})^{1/k} = n^{1/k}d^{1/p}$. For a YES instance, $w_{ij} = t$ for some i, j. Therefore $L_{k,p}(X) \geq t = 2n^{1/k}d^{1/p}$. Thus, there is a constant factor gap in the value. From the known communication complexity lower bounds for multiparty set disjointness [6], [10], [17] for any constant number of passes, the space lower bound for estimating $L_{k,p}$ is $\Omega(nd/t^2) = \Omega(n^{1-2/k}d^{1-2/p})$.

As stated in the introduction, we will design a 1-pass algorithm for estimating $F_{k,p} \triangleq F_k \circ F_p$, $k \geq 1$ and $p \geq 2$. Fix a stream X whose items belong to an arbitrary set $\mathscr{D}$ of size poly(n). We partition items into levels according to their weights and identify levels having a significant contribution to $F_p(X)$.

Notation: Fix an $\eta \geq 1$. We let $\eta = 1 + \gamma$, where $\gamma = \text{poly}(\varepsilon/\log n)$ will be chosen to be sufficiently small. We shall also choose a parameter $\vartheta = \text{poly}(\log(n)/\varepsilon)$ to be large enough, and in particular $\vartheta > \gamma^{-\iota}$ for a large constant $\iota > 1$. We say that *x approximates y within η* if $y \leq x \leq \eta \cdot y$. We say $x \overset{\eta}{=} y$ if $y \leq x \leq (1 + \text{poly}(\gamma))y$.

Definition 3. Choose $\zeta \in [0, \gamma]$ uniformly at random. Define the level sets $S_t(X) = \{a \in \mathscr{D} : |w_a(X)| \in [\zeta\eta^{t-1}, \zeta\eta^t)\}$, for $1 \leq t \leq C$, where $C = O(\log_\eta(\text{poly}(n)/\zeta))$ is the total number of level sets. Let $B \geq 1$ be a parameter. Call a level t *contributing* if $|S_t(X)| \cdot \zeta^p \eta^{pt} \geq \frac{F_p(X)}{B\vartheta}$. For a contributing level t, items in $S_t(X)$ will be called *contributing* items.

2.1. Fuzzy intervals and items

Define the *fuzzy intervals* to be $[\zeta\eta^{t-1}, \zeta\eta^{t-1} + \gamma^3\zeta\eta^{t-1})$ and $[\zeta\eta^t - \gamma^3\zeta\eta^{t-1}, \zeta\eta^t)$. For each $a \in \mathscr{D}$ for which $w_a \stackrel{\text{def}}{=} w_a(X) > 0$ (and hence, $|w_a| \geq 1$), let $\mathscr{E}_a$ be the event that w_a lies in a fuzzy interval, in which case a is *fuzzy*. Let $I(a,t)$ be an indicator which is 1 iff $a \in S_t$. Then $\Pr[\mathscr{E}_a] = \sum_t \Pr[\mathscr{E}_a \mid I(a,t) = 1]\Pr[I(a,t) = 1]$. Let t^* be such that $w_a/\eta^{t^*} < \gamma$ but $w_a/\eta^{t^*-1} \geq \gamma$. Then $I(a,t) = 0$ if $t < t^*$ since $\zeta \leq \gamma$. Also, $\Pr[I(a,t^*) = 1] = \Pr[\zeta\eta^{t^*-1} \leq w_a < \zeta\eta^{t*}] = \Pr[\zeta \geq w_a/\eta^{t*}]$, since $w_a/\eta^{t^*-1} \geq \gamma$. But $\Pr[\zeta \geq w_a/\eta^{t^*}] = \gamma - w_a/\eta^{t^*} = \gamma - w_a/(\eta\eta^{t^*-1}) \leq \gamma - \gamma/\eta \leq \gamma^2$.

For $t > t^*$, to compute $\Pr[\mathscr{E}_a \mid I(a,t) = 1]$, we compute $\Delta_1 = \Pr[\zeta \in [w_a/(\eta^{t-1}(1+\gamma^3)), w_a/\eta^{t-1}]] = (w_a\gamma^3)/(\eta^{t-1}(1+\gamma^3)) \leq w_a\gamma^3/\eta^{t-1}$. Also, $\Delta_2 = \Pr[\zeta \in [w_a/\eta^t, w_a/(\eta^t - \gamma^3\eta^{t-1})]] = (w_a/\eta^{t-1})(1/(\eta - \gamma^3) - 1/\eta) = w_a\gamma^3/(\eta^t(\eta - \gamma^3)) \leq w_a\gamma^3/\eta^{t-1}$. Finally, $\Delta_3 = \Pr[\zeta \in [w_a/\eta^t, w_a/\eta^{t-1}]] = w_a\gamma/\eta^t$. Now, $\Pr[\mathscr{E}_a \mid I(a,t) = 1] = (\Delta_1 + \Delta_2)/\Delta_3 \leq 2\gamma^2\eta \leq 3\gamma^2$. Hence, $\Pr[\mathscr{E}_a] \leq 3\gamma^2 + \gamma^2 = O(\gamma^2)$.

Let X' be the stream of updates restricted to non-fuzzy items. With probability at least $1 - O(\gamma^{1/2})$, $F_p(X') \geq (1 - \gamma^{3/2})F_p(X)$. Also, note that with probability $1 - 1/\text{poly}(n)$, $\zeta \geq 1/\text{poly}(n)$, and so the number of level sets $C = O(\log n)/\gamma$. As $\mathbf{E}[|S_t(X)| - |S_t(X')|] = O(\gamma^2)|S_t(X)|$, $\Pr[|S_t(X)| - |S_t(X')| > \gamma^{1/2}|S_t(X)|] = O(\gamma^{3/2})$, so $\Pr[\exists t \mid |S_t(X)| - |S_t(X')| > \gamma^{1/2}|S_t(X)|] = O(\gamma^{1/2}\log n) = \text{poly}(\gamma)$, for γ sufficiently small. That is,

Lemma 4. *With probability* $1 - \text{poly}(\gamma)$*: for all* t*,* $|S_t(X')| \geq (1 - \text{poly}(\gamma))|S_t(X)|$*, and* $F_p(X) \geq F_p(X') \geq (1 - \text{poly}(\gamma))F_p(X)$.

We condition on the events of Lemma 4.

2.2. An approximate F_p-sampling algorithm

The main result of this section is a sampling algorithm geared towards contributing items.

Theorem 5. *There is a* 1*-pass procedure* $\texttt{Sample}(X, Q; B, \eta)$ *using space* $\widetilde{O}((B^{2/p} + Q^{2/p}) \cdot |\mathscr{D}|^{1-2/p})$ *that outputs the following with probability* $\geq 1 - \text{poly}(\gamma)$*:*

1) *a set* G *including all contributing levels and values* s_t *for* $t \in G$ *with* $s_t \xleftrightarrow{1+\text{poly}(\gamma)} |S_t(X)|$*. If* $t \in G$*, then* $|S_t(X)|\zeta^p\eta^{pt} \geq F_p(X)/(B\vartheta(1 + \text{poly}(\gamma)))$.
2) *A quantity* Φ *such that* $\Phi \stackrel{\eta}{=} F_p(X)$.
3) *For each* $t \in G$*, a set* A_t *formed as follows: either (1)* $S_t(X') \subseteq A_t \subseteq S_t(X)$ *is arbitrary, or (2) a subset* B_t *of* $S_t(X)$ *of size at least* $Q \cdot \frac{\vartheta^2 s_t\zeta^p\eta^{pt}}{\Phi}$ *is chosen at random, and* A_t *is arbitrary subject to* $B'_t \subseteq A_t \subseteq B_t$*, where* B'_t *denotes the non-fuzzy items of* B_t*. Further, the subsets* B_t *are independent for different values of* t.

Proof: The algorithm of [19] defines a contributing level t to be one for which $|S_t(X)|\zeta^p\eta^{pt} \geq F_p(X)/(B\vartheta)$. That algorithm returns values s'_t for all t for which $s'_t \leq \eta|S_t(X)|$, and if t contributes, then $s'_t \geq |S_t(X)|$ (see, e.g., Corollary 20 and Lemma 33 of [26], together with the surrounding discussion and Section 3.5. The bounds here follow by replacing ε with $\text{poly}(\varepsilon/\log n)$ and multiplying the estimates there by $(1+\varepsilon)$ to ensure one-sided error). The algorithm also returns $\tilde{F}_p \stackrel{\eta}{=} F_p(X)$. All of this holds with probability $\geq 1 - 1/\text{poly}(n)$.

Let $\sigma = 1 + \text{poly}(\gamma)$ be such that both $\tilde{F}_p \leq \sigma F_p(X)$ and $s'_t \leq \sigma|S_t(X)|$ for all contributing t. Define $\tau = \tilde{F}_p/(B\vartheta\sigma)$. Put t in G iff $s'_t\zeta^p\eta^{pt} \geq \tau$.

Claim 6. *If* t *is contributing, then* t *is in* G.

Claim 7. *If* t *is in* G*, then* $s'_t \geq |S_t(X)|/\sigma$.

Claim 8. *If* t *is in* G*, then* $|S_t(X)|\zeta^p\eta^{pt} \geq F_p(X)/(B\vartheta\sigma^2)$.

Let $s_t = \sigma s'_t$ for each $t \in G$. Claims 6, 7, and 8 now imply part 1. The space is determined by the algorithm of [19], and is $\widetilde{O}((B\vartheta)^{2/p} \cdot |\mathscr{D}|^{1-2/p}) = \widetilde{O}(B^{2/p} \cdot |\mathscr{D}|^{1-2/p})$. For part 2, let $\Phi \triangleq \sum_{t\in G} s_t\zeta^p\eta^{pt}$.

Claim 9. *For large enough* ϑ*,* $F_p(X) \leq \Phi \leq \sigma^2\eta^p F_p(X)$

For Part 3, fix a $t \in G$ and let $\alpha_t = \frac{s_t\zeta^p\eta^{pt}}{\Phi} \cdot Q$. The quantity α_t represents the expected number of samples that are needed from level t. Assume, w.l.o.g., that $Q \geq 2B\sigma^2\eta^p\vartheta^2$; this will affect the space bound claimed in the theorem by only an $\widetilde{O}(1)$ factor. For $t \in G$, $\alpha_t = \frac{s_t\zeta^p\eta^{pt}}{\Phi} \cdot Q \geq \frac{|S_t(X)|\zeta^p\eta^{pt}}{\sigma^2\eta^p F_p(X)} \cdot Q = \frac{|S_t(X)|\zeta^p\eta^{pt}}{F_p(X)} \cdot \frac{Q}{\sigma^2\eta^p} \geq \frac{F_p(X)}{F_p(X)B\vartheta(1+\text{poly}(\gamma))} \cdot \frac{Q}{\sigma^2\eta^p} \geq \frac{Q}{2B\vartheta\sigma^2\eta^p} \geq \vartheta$, where the first inequality follows from parts 1 and 2, the second inequality by the second property of a $t \in G$ given in part 1, and the last inequality our bound on Q.

We show how to obtain a near-uniform set of roughly $\beta_t = \min(\vartheta^2\alpha_t, s_t)$ samples from each $t \in G$. Let $j \geq 0$ be such that $s_t/2^j \leq \beta_t < s_t/2^{j-1}$.

The key idea is sub-sampling: for each $j \in [\log|\mathscr{D}|]$, let $h_j : \mathscr{D} \mapsto \{0,1\}$ be a random function such that $h_j(a) = 1$ with probability $1/2^j$ and the values $h_j(a)$ for all a are jointly independent. We create a substream Y_j for each j, and items a for which $h_j(a) = 0$ in the corresponding substream are discarded. By Markov's inequality and a union bound, with probability at least $1 - 2/\vartheta$, (*) $F_p(Y_j) \leq (\log|\mathscr{D}|)\vartheta F_p(X)/2^j$ and (**) $L_0(Y_j) \leq (\log|\mathscr{D}|)\vartheta|\mathscr{D}|/2^j$.

Now $\frac{s_t}{2^j} \leq \beta_t \leq \vartheta^2\alpha_t = \frac{\vartheta^2 s_t\zeta^p\eta^{pt}}{\Phi}Q$, which by ap-

plying part 2 yields $\eta^{pt} \geq \frac{\Phi}{\zeta^p \vartheta^2 2^j Q} \geq \frac{F_p}{\zeta^p \vartheta^2 2^j Q}$. So, $\eta^{2t} = (\eta^{pt})^{2/p} \geq \left(\frac{F_p(X)}{\zeta^p \vartheta^2 2^j Q}\right)^{2/p} \geq \left(\frac{F_p(Y_j)}{\zeta^p \vartheta^3 (\log|\mathscr{D}|) Q}\right)^{2/p} \geq \frac{F_2(Y_j)(\vartheta(\log|\mathscr{D}|)|\mathscr{D}|/2^j)^{2/p-1}}{(\zeta^p \vartheta^3 (\log|\mathscr{D}|) Q)^{2/p}} \geq \frac{F_2(Y_j)}{\zeta^2 (\log|\mathscr{D}|)\vartheta^4 Q^{2/p} \cdot |\mathscr{D}|^{1-2/p}}$, where the first inequality follows from our previous calculation, the second from (*), the third from Hölder's inequality, and the last inequality since $p \geq 2$, so in particular $(2^j)^{1-2/p} \geq 1$. Thus by running the F_2-heavy hitters algorithm of [11] on each Y_j, every sub-sampled item of S_t will be returned in 1-pass with space $\widetilde{O}(Q^{2/p}|\mathscr{D}|^{1-2/p})$ with probability $1 - O(\vartheta^{-1})$. However, the algorithm does not know which items it returns are in S_t. Nevertheless, with space $\widetilde{O}(Q^{2/p}|\mathscr{D}|^{1-2/p})$, for each item a returned by the algorithm, it also returns an estimate w'_a for which $w_a \leq w'_a \leq (1+\gamma^3/\eta)\, w_a$.

Now we have two cases. Suppose $\beta_t = s_t$. In this case $j = 0$, and there is no sub-sampling. It follows that every item of S_t is returned by the algorithm of [11]. On the other hand, suppose $\beta_t = \vartheta^2 \alpha_t < s_t$. Then the expected number of items in S_t in Y_j is $\Theta(\vartheta^2 \alpha_t) = \Omega(\vartheta^3)$. It follows by a Chernoff bound that with probability $1 - e^{-\Omega(\vartheta)}$, the number of items in S_t in Y_j will be at least $\vartheta \alpha_t$. Hence, with probability $1 - O(\vartheta^{-1})$, either the entire set S_t, or at least $\vartheta \alpha_t$ samples will be in the output of the algorithm of [11].

Moreover, for every item returned by the F_2-heavy hitters algorithm, we could naively spend an additional pass to determine its exact frequency, and thus which S_t it belongs to. To implement this in 1 pass, the algorithm declares a to be *borderline* if w'_a is in $\cup_t [\zeta\eta^{t-1}, \zeta\eta^{t-1} + \gamma^3\zeta\eta^{t-1})$. If a is borderline, then $w'_a \in [\zeta\eta^{t-1}, \zeta\eta^{t-1} + \gamma^3\zeta\eta^{t-1})$ for some value of t. This means $\zeta\eta^{t-1}/(1+\gamma^3/\eta) \leq w_a < \zeta\eta^{t-1} + \gamma^3\zeta\eta^{t-1}$. But $\zeta\eta^{t-1}/(1+\gamma^3/\eta) \geq \zeta\eta^{t-1} - \gamma^3\zeta\eta^{t-2}$, so a is fuzzy.

If a is not borderline, then using w'_a, a is correctly classified. Indeed, since a is not borderline, $w'_a \in [\zeta\eta^{t-1} + \gamma^3\zeta\eta^{t-1}, \zeta\eta^t)$ for some value of t. As $w_a \leq w'_a$, a misclassification can only occur if $w_a < \zeta\eta^{t-1}$. But $w_a \geq w'_a/(1+\gamma^3/\eta) \geq (\zeta\eta^{t-1} + \gamma^3\zeta\eta^{t-1})/(1+\gamma^3/\eta) \geq \zeta\eta^{t-1} + \gamma^3\zeta\eta^{t-1} - \gamma^3\zeta\eta^{t-2} - \gamma^6\zeta\eta^{t-2} = \zeta\eta^{t-1} + \gamma^4\zeta\eta^{t-2} - \gamma^6\zeta\eta^{t-2} > \zeta\eta^{t-1}$, so a is classified correctly.

This analysis was for a given $t \in G$. For sufficiently large ϑ (as a function of γ), by a union bound over all $O(\log n)/\gamma$ many $t \in G$, the above events occur for all $t \in G$ with probability $\geq 1 - \text{poly}(\gamma)$.

If $\beta_t = s_t$, then the set A_t is S_t minus borderline items. Otherwise, B_t is a random subset of $S_t(X)$ of the desired size, and A_t denotes the items in B_t that were not discarded. As only borderline (and hence fuzzy) items were discarded, this completes the proof. ■

For Section 2.4, we need the following algorithm.

Algorithm 1 Generator$(Q, \{A_t, s_t\}_t, \Phi, G)$

1) Initialize an array T of length Q.
2) For each $i = 1, 2, \ldots, Q$,
   a) Choose a value of t with probability $\frac{\zeta^p \eta^{pt} s_t}{\Phi}$.
   b) Let $T[i]$ be a random item of A_t.
3) Output T.

2.3. A new 1-pass F_k-algorithm

We describe a new 1-pass $\tilde{O}(|\mathscr{D}|^{1-1/k})$-space algorithm OracleEstimator to perform F_k-estimation on a stream X, $k \geq 1$, whose items belong to a set $\mathscr{D}'$ of size poly(n). While the space is inferior, the algorithm will only have limited oracle access to X.

For flexibility, we fix an $\eta' \geq 1$, where $\eta' = 1 + \gamma'$ for $\gamma' = \text{poly}(\varepsilon/\log n)$. It will turn out in our application in Section 2.4 that η' will be slightly larger than η. In this section we will use the same sufficiently large parameter $\vartheta = \text{poly}(\log(n)/\varepsilon)$ as in the previous section. Choose $\zeta' \in [0, \gamma']$ uniformly at random. Define the level sets $S_t(X) = \{a \in \mathscr{D}' : |w_a(X)| \in [\zeta'(\eta')^{t-1}, \zeta'(\eta')^t)\}$, for $1 \leq t \leq C'$, where $C' = O(\log_{\eta'}(\text{poly}(n)/\zeta'))$ is the total number of level sets. We condition on $C' = O(\log n)/\gamma'$, which occurs with probability $1 - 1/\text{poly}(n)$. Notice that we assume, if $w_a(X) \neq 0$, then $|w_a(X)| \geq 1$.

Call a level t *important* if $|S_t(X)| \cdot (\zeta')^k (\eta')^{kt} \geq \frac{F_k(X)}{\vartheta}$. For an important level t, items in $S_t(X)$ will be called *important* items. The algorithm is allowed to specify an integer $j \in [C']$, and a value Q. The oracle forms a sub-stream by including each item in $\mathscr{D}'$ independently with probability 2^{-j}. Call the included items the *survivors*, and the set of such survivors S. The oracle returns Q i.i.d. samples from a distribution on S with the following two properties: (1) survivor a is returned with probability $(1 \pm O((\gamma')^3/\eta')) \frac{|w_a(X)|}{\sum_{b \in S} |w_b(X)|} \pm \frac{2}{\vartheta Q}$, (2) if $w_a(X) = 0$, then survivor a is returned with probability 0. The oracle also returns a $(1 + O((\gamma')^3/\eta'))$-approximation to the L_1-norm of the vector of survivors (using, say, the algorithm of [21]), which can be done in $\text{poly}(\varepsilon^{-1}\log n)$ space. Notice that $\sum_{\text{unimportant } t} |S_t(X)|(\zeta')^k(\eta')^{tk} \leq \gamma' \cdot F_k(X)$, provided $\vartheta = \text{poly}(\varepsilon^{-1}\log n)$ is large enough, and so $\sum_{\text{important } t} |S_t(X)|(\zeta')^k(\eta')^{tk} \geq (1-\gamma') F_k(X)$.

Theorem 10. *Given the above oracle, there exists an algorithm* OracleEstimator *taking* $\tilde{O}(|\mathscr{D}'|^{1-1/k})$ *non-adaptive samples, and outputting a* $(1 \pm \text{poly}(\gamma'))$*-approximation to* F_k *with probability* $\geq 1 - \text{poly}(\gamma')$.

Proof: Suppose we obtain approximations s_t to $|S_t|$ for each t with the following two properties:

(1) if S_t is important, then $(1-\text{poly}(\gamma'))|S_t| \le s_t$, and (2) for every t, $s_t \le (1+\text{poly}(\gamma'))|S_t|$. Consider the following estimate $E \triangleq \sum_t s_t(\zeta')^k(\eta')^{kt}$. Then $E \le (\eta')^k(1+\text{poly}(\gamma'))\sum_t |S_t|(\zeta')^k(\eta')^{k(t-1)} \le (\eta')^k(1+\text{poly}(\gamma'))F_k \le (1+\text{poly}(\gamma'))F_k$. On the other hand, $E \ge \sum_{\text{important } t} s_t(\zeta')^k(\eta')^{kt} \ge \sum_{\text{important } t}(1-\text{poly}(\gamma'))|S_t|(\zeta')^k(\eta')^{kt} \ge (1-\text{poly}(\gamma'))F_k$. It suffices to compute the values s_t.

Queries: Put $L = O((\gamma')^{-2}\log|\mathscr{D}'|)$. For each $j \in [\log|\mathscr{D}'|]$ and $\ell \in [L]$ the algorithm asks for an array $T^{j,\ell}$ of $Q = \vartheta^4|\mathscr{D}'|^{1-1/k}$ samples from the oracle. So the number of samples is as claimed.

Algorithm: Let $S^{j,\ell}$ be the set of survivors in the ℓ-th independent sub-sampling phase at rate 2^{-j}.

For each j,ℓ, the algorithm is given the array $T^{j,\ell}$ as well as a $(1\pm O((\gamma')^3/\eta'))$-approximation $h^{j,\ell}$ to the L_1-norm of the stream restricted to items in $S^{j,\ell}$. Let $R^{j,\ell}$ be the set of coordinates which are sampled at least ϑ times in $T^{j,\ell}$. For each $i \in S^{j,\ell}$, let $a_i^{j,\ell}$ be the number of times item i was sampled in $T^{j,\ell}$. Put $b_i^{j,\ell} = \left(\frac{h^{j,\ell}}{Q}\right)\cdot a_i^{j,\ell}$. Then we have $\mathbf{E}[b_i^{j,\ell}] = \mathbf{E}\left[\frac{h^{j,\ell}}{Q}\cdot a_i^{j,\ell}\right] = (1\pm O((\gamma')^3/\eta'))\left(\frac{h^{j,\ell}}{Q}\right)\frac{Q|w_i(X)|}{L_1(S^{j,\ell})} \pm \frac{2Q}{\vartheta Q} = (1\pm O((\gamma')^3/\eta'))|w_i(X)| \pm \frac{2}{\vartheta}$. Since if $w_i(X)=0$ then i is not sampled, i.e., does not occur in any $T^{j,\ell}$, we have $|w_i(X)| \ge 1$. But provided ϑ is large enough, it follows that $\mathbf{E}[b_i^{j,\ell}] = (1\pm O((\gamma')^3/\eta'))|w_i(X)|$.

For all $i \in S^{j,\ell}$, $a_i^{j,\ell}$ is a sum of Q i.i.d. indicator random variables. By Chernoff bounds there is a constant $\alpha > 0$ so that the following events hold with probability $\ge 1-1/\text{poly}(n)$: for all i, $\mathbf{E}[a_i^{j,\ell}] \le \alpha(\eta')^2(\gamma')^{-6}\log|\mathscr{D}'| \Rightarrow a_i^{j,\ell} = O((\eta')^2(\gamma')^{-6}\log|\mathscr{D}'|)$, and $\mathbf{E}[a_i^{j,\ell}] > \alpha(\eta')^2(\gamma')^{-6}\log|\mathscr{D}'| \Rightarrow |a_i^{j,\ell} - \mathbf{E}[a_i^{j,\ell}]| = O((\gamma')^3/\eta')\mathbf{E}[a_i^{j,\ell}]$. By choosing $\vartheta = \omega((\eta')^2(\gamma')^{-6}\log|\mathscr{D}'|)$, these events imply that for all $i \in R^{j,\ell}$, $|a_i^{j,\ell} - \mathbf{E}[a_i^{j,\ell}]| = O((\gamma')^3/\eta')\mathbf{E}[a_i^{j,\ell}]$. Hence, for all $i \in R^{j,\ell}$, $b_i^{j,\ell} = (1\pm O((\gamma')^3/\eta'))|w_i(X)|$. By rescaling and adjusting constants in the $O(\cdot)$, we can ensure that $|w_i(X)| \le b_i^{j,\ell} \le (1+(\gamma')^3/\eta')|w_i(X)|$. The algorithm classifies i as borderline if $b_i^{j,\ell} \in \cup_t[\zeta'(\eta')^{t-1}, \zeta'(\eta')^{t-1} + (\gamma')^3(\eta')^{t-1})$. The algorithm throws away sampled items that are borderline. We adapt the definition of fuzzy intervals in Section 2.1 to ζ', η', γ', as well as Lemma 4 to conclude that with probability $1-\text{poly}(\gamma')$, for all t, $|S_t(X')| \ge (1-\text{poly}(\gamma'))|S_t(X)|$, and $F_k(X) \ge F_k(X') \ge (1-\text{poly}(\gamma'))F_k(X)$. As in the proof of Theorem 5, if a is borderline, then a is fuzzy. Also, if a is not borderline, then a is correctly classified.

For each $t \in [C']$, the algorithm attempts to find a $j(t)$ for which more than a 1/3 fraction of the different values of ℓ, $R^{j(t),\ell}$ contains an item i for which $b_i^{j(t),\ell} \in [\zeta'(\eta')^{t-1}, \zeta'(\eta')^t)$, after discarding borderline items. If there are multiple such j, the algorithm chooses the smallest one. If the algorithm finds such a value $j(t)$, for each $\ell \in [L]$, let $z^{j(t),\ell,t}$ be the number of distinct items i in $R^{j(t),\ell}$ for which $b_i^{j(t),\ell} \in [\zeta'(\eta')^{t-1}, \zeta'(\eta')^t)$ (after discarding borderline items).

If the algorithm finds such a $j(t)$, it sets $s_t = \frac{2^{j(t)}}{L}\sum_{\ell=1}^L z^{j(t),\ell,t}$, otherwise it sets $s_t = 0$. Now the algorithm outputs the estimate E described earlier.

Analysis: For each value of j and ℓ, the number $n^{j,\ell,t}$ of items in $S^{j,\ell}$ from S_t has expectation $|S_t|/2^j$. Since $n^{j,\ell,t}$ is a sum of i.i.d. Bernoulli random variables, and since the $n^{j,\ell,t}$ are jointly independent as we vary ℓ, $\sum_{\ell=1}^L n^{j,\ell,t}$ is a sum of i.i.d. Bernoulli variables, and so by a Chernoff bound the following events hold with probability $\ge 1-1/\text{poly}(n)$: $\forall t, j$ for which $|S_t|2^{-j} > 1/4$, $(1-\gamma')|S_t| \le \frac{2^j}{L}\sum_{\ell=1}^L n^{j,\ell,t} \le (1+\gamma')|S_t|$. Here we use that $L = O((\gamma')^{-2}\log|\mathscr{D}'|)$. We condition on this event in the remainder of the proof. We are only conditioning on the randomness of the sub-sampling, rather than on what the oracle samples from the substreams.

We also condition on the event that the L_1-norm of the survivors, denoted $L_1(S^{j,\ell})$, for each j and ℓ satisfies $L_1(S^{j,\ell}) \le \text{poly}((\gamma')^{-1})(L\log|\mathscr{D}'|)L_1(X)2^{-j}$. This occurs by a Markov and a union bound (over $j,\ell \in [\log|\mathscr{D}'|]$) with probability $\ge 1-\text{poly}(\gamma')$.

Lemma 11. *With probability $\ge 1-1/\text{poly}(n)$, if a value $j(t)$ is found for a given t, then $|S_t|2^{-j(t)} > 1/4$.*

This lemma shows that $\frac{2^{j(t)}}{L}\sum_{\ell=1}^L n^{j(t),\ell,t}$ is a $(1\pm\gamma')$-approximation to $|S_t|$ if a value $j(t)$ is found. Since $z^{j(t),\ell,t} \le n^{j(t),\ell,t}$ for every ℓ and t, and since $s_t = 0$ if no $j(t)$ is found, for every t, $s_t \le (1+\gamma')|S_t|$.

Lemma 12. *With probability $\ge 1-\text{poly}(\gamma')$, $\forall$ important t, a value $j(t)$ is found and $s_t \ge (1-\text{poly}(\gamma'))|S_t|$.*

Proof: If t is important then $|S_t|(\zeta')^k(\eta')^{tk} \ge \frac{F_k}{\vartheta}$, so $|S_t|^{1/k}\zeta'(\eta')^t \ge \frac{F_k^{1/k}}{\vartheta^{1/k}} \ge \frac{L_1(X)}{|\mathscr{D}'|^{1-1/k}\vartheta^{1/k}}$, where the last inequality follows by Hölder's inequality. Consider any value of j for which $1/4 < 2^{-j}|S_t| \le 1$, and fix any $\ell \in [L]$. As conditioned above, $L_1(S^{j,\ell}) \le \text{poly}((\gamma')^{-1})(L\log|\mathscr{D}'|)L_1(X)2^{-j}$. Also, since $k \ge 1$, we have $|S_t|\zeta'(\eta')^t \ge \frac{2^jL_1(S^{j,\ell})}{|\mathscr{D}'|^{1-1/k}(L\log|\mathscr{D}'|)\text{poly}((\gamma')^{-1})\vartheta} \ge \frac{2^jL_1(S^{j,\ell})}{|\mathscr{D}'|^{1-1/k}\vartheta^2}$, since $L\log|\mathscr{D}'|\text{poly}((\gamma')^{-1})\vartheta \le \vartheta^2$.

Using our choice of j, it follows that $\zeta'(\eta')^{t-1} \ge \frac{L_1(S^{j,\ell})}{|\mathscr{D}'|^{1-1/k}\vartheta^2\eta'} = \Omega\left(\frac{L_1(S^{j,\ell})}{|\mathscr{D}'|^{1-1/k}\vartheta^2}\right)$. Since we take $\vartheta^4|\mathscr{D}|^{1-1/k}$ samples from the oracle, by a Chernoff bound, with probability $\ge 1-e^{-\Omega(\vartheta)}$, for every ℓ, $T^{j,\ell}$ will contain

at least ϑ samples from every element of S_t in $S^{j,\ell}$, and so all such elements will be in $R^{j,\ell}$.

By Lemma 4, at most a poly(γ') fraction of S_t is fuzzy. Since $L = \Omega(\log|\mathscr{D}'|)$ and $2^{-j}|S_t| = \Omega(1)$, it follows by a Chernoff bound that with probability $\geq 1 - 1/\mathrm{poly}(n)$, $(1-\mathrm{poly}(\gamma'))\sum_{\ell=1}^{L} n^{j,\ell,t} \leq \sum_{\ell=1}^{L} z^{j,\ell,t}$. It follows that s_t is a $(1\pm\mathrm{poly}(\gamma'))$-approximation to $|S_t|$. By Lemma 12, if a value $j(t)$ is found, then $|S_t|2^{-j(t)} > 1/4$, and in this case s_t $(1\pm\mathrm{poly}(\gamma'))$-approximates $|S_t|$.

To show a value $j(t)$ for which $2^{-j(t)}|S_t| \leq 1$ is found, consider j for which $1/2 < 2^{-j}|S_t| \leq 1$. By a Chernoff bound, since $L = \Omega(\log|\mathscr{D}'|)$, with probability $\geq 1 - 1/\mathrm{poly}(n)$ for more than 5/12 of the values ℓ, $S^{j,\ell}$ contains an element of S_t, and since $2^{-j}|S_t| > 1/4$, for every ℓ, $R^{j,\ell}$ contains an element of S_t if $S^{j,\ell}$ does. After discarding borderline items, with probability $\geq 1 - 1/\mathrm{poly}(n)$, the sampling algorithm still finds at least a $(5/12)(1-\mathrm{poly}(\gamma')) > 1/3$ fraction of different ℓ containing an item in S_t. Hence, a value $j(t)$ is found, and the algorithm chooses the smallest j found. As the above events either occur with probability $\geq 1 - 1/\mathrm{poly}(n)$, or occur simultaneously for all t with probability $\geq 1-\mathrm{poly}(\gamma')$, by a union bound the above events jointly occur for all t with probability $\geq 1-\mathrm{poly}(\gamma')$ (note that the number of t depends on γ', which is why we needed probability of occurrence $\geq 1-1/\mathrm{poly}(n)$ for events pertaining to individual t). ■

The theorem follows. ■

2.4. 1-pass estimation of $F_{k,p}$

We are given a stream X of items of length m, each belonging to $[n]\times[d]$. Let X_i denote the sub-stream of X corresponding to updates to item (i,j) for all $j \in [d]$. We show how to estimate $F_{k,p}(X) = \sum_i\left(\sum_j |w_{ij}(X)|^p\right)^k = \sum_i |F_p(X_i)|^k$. Consider the pseudo-code shown in Algorithm 2 which uses space $\widetilde{O}(n^{1-2/(kp)}d^{1-2/p})$. We now prove correctness.

A few natural events. Let $W^{j,\ell}$ be the stream with all updates to entries that are fuzzy (with respect to ζ,η,γ) removed. Clearly, $F_{k,p}(W^{j,\ell}) \leq F_{k,p}(X^{j,\ell})$ for all j,ℓ. In the sequel, recall $\gamma = \mathrm{poly}(\varepsilon/\log n)$ is sufficiently small.

Lemma 13. *For all $j \in [\log n]$ and $\ell \in [L]$, $\Pr_\zeta\left[\frac{F_{k,p}(W^{j,\ell})}{1-\mathrm{poly}(\gamma)} \geq F_{k,p}(X^{j,\ell})\right] \geq 1-\mathrm{poly}(\gamma)$.*

We condition on some events that jointly occur with probability $\geq 1-\mathrm{poly}(\gamma)$. We condition on the event of Lemma 13, as well as all invocations of Sample succeeding (i.e., meeting the properties of Theorem 5). We also condition on the events of Lemma 4, for all $j \in [\log n]$ and $\ell \in [L]$. We also condition on the event $\mathscr{E}$ that for all $t \in [C]$, $j \in [\log n]$, and $\ell \in [L]$, $|A_t^{j,\ell}| \geq (1-\mathrm{poly}(\gamma))|B_t^{j,\ell}|$. As each $\mathbf{E}[|A_t^{j,\ell}|]$ is at least the size of $B_t^{j,\ell}$ with all fuzzy items removed, it is at least $(1-O(\gamma^2))|B_t^{j,\ell}|$, so by a union bound over the $O(\log n)/\gamma$ many t, $j \in [\log n]$, and $\ell \in [L]$, a Markov bound shows $\mathscr{E}$ occurs with probability $\geq 1-\mathrm{poly}(\gamma)$.

Algorithm 2 Compute $F_{k,p}(X)$

1) For each $j \in [\log n]$ and each $\ell \in [L]$, where $L = O((\gamma')^{-2}\log n)$,
   a) Keep each row of the matrix with probability 2^{-j}. Let $X^{j,\ell}$ be the restriction of X to updates to the set $S^{j,\ell}$ of surviving rows.
   b) Call $\mathtt{Sample}(X^{j,\ell}, Q; B, \eta)$ with $B = \vartheta^5 n^{1-1/k}$, $Q = 2\vartheta^4 n^{1-1/k}$, and parameter p to obtain the $A_t^{j,\ell}, s_t^{j,\ell}, \Phi^{j,\ell}$, and $G^{j,\ell}$, together with sets $Z^{j,\ell}$ of items declared borderline.
2) Put $Z = \cup_{j,\ell} Z^{j,\ell}$.
3) For each $j \in [\log n]$ and $\ell \in [L]$,
   a) For each $t \in [C]$, set $A_t^{j,\ell} = A_t^{j,\ell} \setminus Z$.
   b) $T^{j,\ell} = \mathsf{Generator}(Q, \{A_t^{j,\ell}, s_t^{j,\ell}\}_t, \Phi^{j,\ell}, G^{j,\ell})$.
4) $\forall j,\ell$, replace pairs $(a,b) \in T^{j,\ell}$ with row IDs a.
5) Feed the $T^{j,\ell}$ and $\Phi^{j,\ell}$ (here $h^{j,\ell}$ in Section 2.3 is set to $\Phi^{j,\ell}$) for all j,ℓ into OracleEstimator to estimate F_k, and output its output.

We further define and condition on the following event $\mathscr{F}$. Say a row a in X is *obscured* if $F_p(X_a) \geq (1 + O((\gamma')^3/\eta'))F_p(W_a)$. Then, $\Pr_\zeta[a \text{ is obscured}] \leq \mathrm{poly}(\gamma)$, and so with probability $\geq 1-\mathrm{poly}(\gamma)$, at most $n\mathrm{poly}(\gamma)$ rows are obscured. Moreover, $\sum_{\text{obscured } a}|F_p(X_a)|^k \leq \mathrm{poly}(\gamma)F_{k,p}(X)$ with probability $\geq 1-\mathrm{poly}(\gamma)$. The event $\mathscr{F}$ is the joint occurrence of these two events.

Let V be the matrix obtained by deleting the set Z of borderline items from X, and let $V^{j,\ell}$ be the matrix obtained by deleting the items in Z from $X^{j,\ell}$. As any borderline item (as classified by any invocation of Sample) is also fuzzy, $(1-\mathrm{poly}(\gamma))F_{k,p}(X^{j,\ell}) \leq F_{k,p}(W^{j,\ell}) \leq F_{k,p}(V^{j,\ell}) \leq F_{k,p}(X^{j,\ell})$.

Lemma 14. *Suppose $\gamma \leq (\gamma')^q$ for a sufficiently large constant $q > 0$. Then with probability $\geq 1-\mathrm{poly}(\gamma)$, for all j,ℓ, $F_p(V^{j,\ell}) \leq \Phi^{j,\ell} \leq (1 \pm O((\gamma')^3/\eta'))F_p(V^{j,\ell})$, where the constant in the $O(\cdot)$ can be made arbitrarily small (here the matrix structure of $V^{j,\ell}$ is ignored).*

Overall strategy. We will show that for every row $a \in [n]$ that is not obscured, for every $j \in [\log n]$ and $\ell \in [L]$, the i-th sample of $T^{j,\ell}$ from a surviving row a is $(1\pm\mathrm{poly}(\gamma))\frac{F_p(V_a)}{F_p(V^{j,\ell})} \pm \frac{2}{Q\vartheta}$. Provided that γ is sufficiently small, this probability is $(1\pm\Theta((\gamma')^3/\eta'))\frac{F_p(V_a)}{F_p(V^{j,\ell})} \pm \frac{2}{Q\vartheta}$.

If all rows were not obscured, then, by Theorem 10, with probability $\geq 1-\text{poly}(\gamma)$, the output of Algorithm 2 will be a $(1\pm\text{poly}(\gamma))$-approximation to $F_{k,p}(V)$, and hence $F_{k,p}(X)$, completing the proof. It turns out that obscured rows do not pose much of a problem, and can be handled by slight modifications to OracleEstimator. We first consider unobscured rows a, then explain the modifications to OracleEstimator.

Fix $j\in[\log n]$ and $\ell\in[L]$. In Generator, $\Pr[T^{j,\ell}[i]=a]=\sum_{t\in G^{j,\ell}}\frac{\zeta^p\eta^{pt}s_t^{j,\ell}}{\Phi^{j,\ell}}\cdot\frac{|\{b|(a,b)\in A_t^{j,\ell}\}|}{|A_t^{j,\ell}|}$. There are two cases for a given $t\in G^{j,\ell}$, depending on the property of $A_t^{j,\ell}$ given in step 3 of Theorem 5.

Finishing the analysis: In the full version we show that no matter which case occurs, $\Pr[T^{j,\ell}[i]=a]=\sum_{t\in G^{j,\ell}}\frac{\zeta^p\eta^{pt}s_t^{j,\ell}}{\Phi^{j,\ell}}\cdot\frac{|\{b|(a,b)\in A_t^{j,\ell}\}|}{|A_t^{j,\ell}|}$ is at least $(1-\text{poly}(\gamma))\cdot\left(\frac{F_p(W_a)}{F_p(V^{j,\ell})}-\frac{2}{Q\vartheta}\right)\geq(1-\text{poly}(\gamma))\cdot\frac{F_p(W_a)}{F_p(V^{j,\ell})}-\frac{2}{Q\vartheta}$, and at most $(1+\text{poly}(\gamma))\frac{F_p(X_a)}{F_p(V^{j,\ell})}+\frac{1}{2Q\vartheta}$. Since a is not obscured, $F_p(X_a)\leq(1+O((\gamma')^3/\eta'))F_p(W_a)$. As $F_p(W_a)\leq F_p(V_a)\leq F_p(X_a)$, it follows that for small enough γ, $\Pr[T^{j,\ell}[i]=a]=(1\pm O((\gamma')^3/\eta'))\frac{F_p(V_a)}{F_p(V^{j,\ell})}\pm\frac{2}{Q\vartheta}$.

The only remaining issue is that obscured rows a need not satisfy $\Pr[T^{j,\ell}[i]=a]=(1\pm O((\gamma')^3/\eta'))\frac{F_p(V_a)}{F_p(V^{j,\ell})}$. However, we can slightly adapt the proof of Theorem 5 to handle this. In that proof, we define the $a_i^{j,\ell}$ and $b_i^{j,\ell}$ variables as before, and similarly deduce that for all j,ℓ, $|b_i^{j,\ell}-\mathbf{E}[b_i^{j,\ell}]|=O((\gamma')^3/\eta')\mathbf{E}[b_i^{j,\ell}]$. If row i is not obscured, then $b_i^{j,\ell}=(1\pm O((\gamma')^3/\eta'))F_p(V_i)$. However, if row i is obscured, our bounds above only guarantee $(1-O((\gamma')^3/\eta'))F_p(W_i)\leq b_i^{j,\ell}\leq(1+O((\gamma')^3/\eta'))F_p(X_i)$. OracleEstimator then classifies row i in each sub-sampling experiment (j,ℓ). If in any two experiments the classification differs, or i is ever classified as borderline (w.r.t. ζ',η',γ'), then row i is dropped. Rows that are not obscured do not have differing classifications, and because of event $\mathscr{F}$, $\sum_{\text{obscured i}}|F_p(X_i)^k\leq\text{poly}(\gamma)F_{k,p}(V)$, so dropping all such rows only changes the output of OracleEstimator by a $(1-\text{poly}(\gamma))$ factor. It may happen that an obscured row is not dropped, but then since $(1-O((\gamma')^3/\eta'))F_p(W_i)\leq b_i^{j,\ell}\leq(1+O((\gamma')^3/\eta'))F_p(X_i)$, it follows that OracleEstimator approximates the F_k-value of a vector v, where $v_i=F_p(V_i)$ if i is not obscured, and $v_i\in[0,F_p(X_i)]$ if i is obscured. By the above, $F_k(v)=(1\pm\text{poly}(\gamma))F_{k,v}(X)$, as desired. One final issue is that after a subset of the obscured rows are dropped, OracleEstimator still needs $\vartheta^4n^{1-1/k}$ samples for each j,ℓ in the proof of Theorem 5. This still holds because $\mathbf{E}[\sum_{\text{obscured }i}F_p(X_i^{j,\ell})]\leq\text{poly}(\gamma)F_p(X^{j,\ell})$, so by a Markov bound and a union bound, with probability $\geq 1-\text{poly}(\gamma)$, for all j,ℓ, $\sum_{\text{obscured }i}F_p(X_i^{j,\ell})\leq\text{poly}(\gamma)F_p(X^{j,\ell})\leq\text{poly}(\gamma)F_p(V^{j,\ell})$. Because $Q=2\vartheta^4n^{1-1/k}$ and the probability any entry of $T^{j,\ell}$ is obscured is bounded by $\text{poly}(\gamma)$, it follows by a Chernoff bound that w.h.p., for all j,ℓ, there are $\vartheta^4n^{1-1/k}$ entries that are not dropped.

3. An Optimal Lower Bound for $L_{k,0}$ and $L_{k,1}$

We prove an $\Omega(n^{1-1/k})$ space bound for estimating $L_{k,p}$ where $k>0$ and $p=0,1$. This is achieved via the two-party communication problem defined below.

Let $H=\{0,1\}^d$ denote the Hamming cube of dimension d with distance $|\cdot|$. In the communication problem f, Alice gets an input $x=(x_1,x_2,\ldots,x_n)\in H^n$ and Bob gets an input $(y_1,y_2,\ldots,y_n)\in H^n$. The No instances satisfy the promise that $|x_i-y_i|\leq 1$ for all i. The Yes instances satisfy the promise that there is a coordinate j such that $|x_i-y_i|=d$ for some coordinate j. Thus, $f(x,y)=\bigvee_i g(x_i,y_i)$ where $g(u,v)=1$ if $|u-v|=d$ and $g(u,v)=0$ if $|u-v|\leq 1$. This property suggests a direct-sum argument for the communication complexity of f.

We first check that f yields a space lower bound for estimating $L_{k,p}$ in a data stream. Interpret the bits of x as positive 0/1 updates and the bits of y as negative 0/1 updates and concatenate to form an single input $x\circ y$. Then $L_{k,p}(x\circ y)=(\sum_i|x_i-y_i|^k)^{1/k}$. This is at most $n^{1/k}$ for a No instance but at least d for a Yes instance. Set $d=2n^{1/k}$ to yield a constant factor gap between the two values. Below, we will prove an $\Omega(n/d)$ communication lower bound for f.

We now briefly review the information complexity paradigm for proving communication lower bounds via direct sum arguments, as developed in [6], for two-party communication protocols. Let μ be a distribution on the inputs (X,Y) of Alice and Bob, denoted by $(X,Y)\sim\mu$. We say that μ is *product* if X and Y are *independent*. Non-product distributions are handled via an auxiliary random variable $D\sim\nu$ such that X and Y are independent conditioned on D, denoted by $X,Y\perp D$. Given a randomized private-coin protocol Π, let $\Pi(x,y)$ be a random variable denoting the transcript of the communication between Alice and Bob on inputs x and y. Consider the joint probability space on X, Y, D, and the randomness used by the players such that (1) the joint distribution of (X,Y,D) and the randomness are independent (2) $(X,Y)\sim\mu$, (3) $D\sim\nu$, and (4) $X,Y\perp D$. The *(conditional) information cost* of Π under (μ,ν) is defined to be $I(X,Y:\Pi(X,Y)\mid D)$, where $(X,Y)\sim\mu,D\sim\nu$. Since $I(X,Y:\Pi(X,Y)\mid D)\leq$

$H(\Pi(X,Y)) \le |\Pi|$, it suffices to prove lower bounds on the information cost of a correct protocol. Throughout this section, we define a correct protocol to be one whose error probability is a sufficiently small constant.

For the communication problem $f = \bigvee_i g$ given above, we first define a distribution (μ', ν') for $g(u,v)$ where $(U,V) \sim \mu'$ and the auxiliary random variable is a pair $(S,T) \sim \nu'$: $S \in_R \{\text{A},\text{B}\}$ and T is a uniformly chosen pair (u,v) such that $u \in_R H$ and v is chosen uniformly from the neighbors of u in H. If $S = \text{A}$, then $U \in_R \{u,v\}$ and $V = v$. Otherwise $S = \text{B}$, and here $U = u$ and $V \in_R \{u,v\}$. Note that $U \perp V \mid S,T$.

The distribution for $f(x,y)$ is defined by letting $\mu = \mu'^n$ and $\nu = \nu'^n$. In other words, if $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$ are the inputs and $((S_1,T_1),(S_2,T_2),\ldots,(S_n,T_n))$ are the auxiliary random variables, then independently for each i, $(X_i,Y_i) \sim \mu'$ and $(S_i,T_i) \sim \nu'$.

Proposition 15 (Direct Sum[6])**.** *Let* $\text{IC}_\mu(f \mid \nu)$ *(resp.* $\text{IC}_{\mu'}(g \mid \nu')$*) denote the minimum information cost of a correct protocol computing* f *(resp.* g*) under* (μ,ν) *(resp.* (μ',ν')*). Then* $CC(f) \ge \text{IC}_\mu(f \mid \nu) \ge n \cdot \text{IC}_{\mu'}(g \mid \nu')$.

Via the direct sum theorem, the following implies our communication complexity lower bound for f.

Theorem 16. $\text{IC}_{\mu'}(g \mid \nu') = \Omega(1/d)$.

Proof: Let Π be a correct protocol for g and let $\pi_{u,v}$ denote the probability distribution over transcripts induced by Π on input (u,v). Let $\psi_{u,v} \in \ell_2$ be obtained via the square-root map $\psi_{u,v}(\tau) = \sqrt{\pi_{u,v}(\tau)}$ for all transcripts τ. Note that $\|\psi_{u,v}\| = 1$ where $\|.\|$ denotes the standard ℓ_2 norm. Following [20], $\psi_{u,v}$ is called the *transcript wave function* of (u,v) in Π.

The *Hellinger distance* between two transcript wave functions $\psi_{u,v}$, $\psi_{u',v'}$ is a scaled Euclidean distance equal to $\frac{1}{\sqrt{2}}\|\psi_{u,v} - \psi_{u',v'}\|$.

Proposition 17 ([6])**.** *Let* $\|\cdot\|' \triangleq \frac{1}{2}\|\cdot\|^2$.
Information-to-Hellinger: *If* $(U,V) \in_R \{(u,v),(u',v')\}$ *then* $I(U,V : \Pi(U,V)) \ge \|\psi_{u,v} - \psi_{u',v'}\|'$.
Soundness: $\|\psi_{u,v} - \psi_{u',v'}\|' = \Omega(1)$ *if* $g(u,v) \ne g(u',v')$
Pythagorean property: $\|\psi_{u,v} - \psi_{u',v'}\|' \ge \frac{1}{2} \cdot (\|\psi_{u,v} - \psi_{u,v'}\|' + \|\psi_{u',v} - \psi_{u',v'}\|')$

We now bound the information cost of Π:

$$
\begin{aligned}
&I(U,V : \Pi(U,V) \mid S,T) \\
&= \mathbb{E}_{s \in \{\text{A},\text{B}\}, u \in H, j \in [d]}\, I(U,V : \Pi(U,V) \mid S = s, T = (u, u+e_j)) \\
&= \tfrac{1}{2} \cdot \mathbb{E}_{u \in H, j \in [d]}\, I(V : \Pi(u,V) \mid T = (u, u+e_j)) \\
&\quad + I(U : \Pi(U, u+e_j) \mid T = (u, u+e_j)) \\
&\ge \tfrac{1}{2} \cdot \mathbb{E}_{u \in H, j \in [d]} \|\psi_{u,u} - \psi_{u,u+e_j}\|' + \|\psi_{u,u+e_j} - \psi_{u+e_j,u+e_j}\|', \qquad (1)
\end{aligned}
$$

where the last inequality follows by relating mutual information to Hellinger distance. Applying the triangle inequality to the expression within the expectation (which incurs a loss of 1/2 due to squared Euclidean distances), the above quantity is at least

$$\tfrac{1}{4} \cdot \mathbb{E}_{u \in H, j \in [d]} \|\psi_{u,u} - \pi_{u+e_j,u+e_j}\|' \qquad (1)$$

The short diagonals property (see [22]) implies that for any family $\{\rho_u\}_{u \in H}$ of elements in ℓ_2, $\mathbb{E}_{u \in H, j \in [d]} \|\rho_u - \rho_{u+e_j}\|' \ge \frac{1}{d} \cdot \mathbb{E}_{u \in H} \|\rho_u - \rho_{\overline{u}}\|'$, where $\overline{u}$ is the bit-wise complement of u. Applying the above bound with $\rho_u = \psi_{u,u}$ in (1), the information cost is at least $\frac{1}{4d} \cdot \mathbb{E}_{u \in H} \|\psi_{u,u} - \psi_{\overline{u},\overline{u}}\|' \ge \frac{1}{8d} \cdot \mathbb{E}_{u \in H} \|\psi_{u,u} - \psi_{u,\overline{u}}\|' + \|\pi_{\overline{u},u}, \pi_{\overline{u},\overline{u}}\|'$, where the last inequality follows from the Pythagorean property. Now, since each Hellinger distance expression involves a YES-instance and a NO-instance, the soundness property implies that the information cost is $\Omega(1/d)$. ■

4. Solution Sketches to Other Problems

Estimating $L_{k,2}$ for $k > 0$: The main idea is to use an existing L_k-approximation algorithm to estimate $L_k(u)$, where u is a random variable satisfying $\mathbb{E}[L_k(u)] = \mu_k L_{k,2}(X)$ and $\mathbf{Var}[L_k(u)] = \mu_{2k} L_{2k,2}(X)$, where μ_k and μ_{2k} are constants depending only on k, and reduce the variance by averaging several such estimators.

To define u, we (pseudo-) randomly choose n vectors v_i of d-dimensional independent normal variables, and set $u_i = \langle v_i, X_i \rangle$. Using the k-th moments of the half-normal distribution and the 2-stability of the normal distribution, we show $\mathbb{E}[|u_i|^k] = \mu_k \|X_i\|_2^k$ and establish the bounds on $\mathbb{E}[L_k(u)]$ and $\mathbf{Var}[L_k(u)]$ given above. Hence, the space is the same as that of estimating L_k, up to a $\Theta(\varepsilon^{-2})$ factor.

Finding heavy rows according to L_2: This is denoted as $Freq(L_2)$ in [13]. Given $\phi > 0$, let $W = L_{1,2}(X) = \sum_i \|X_i\|_2$ and define the set $HH_\phi(X) = \{i \mid \|X_i\|_2 \ge \phi W\}$. The problem is to return a set T for which $HH_\phi(X) \subseteq T \subseteq HH_{\phi-\epsilon}(X)$. We solve the related point query problem, i.e. estimate $\|X_i\|_2$ for all rows i within an additive error of γW, for a given $\gamma > 0$. With $\gamma = \epsilon/2$, and using an approximation to W, we solve $Freq(L_2)$ by including those rows we estimate as being heavy. Let u be as in the previous algorithm for $L_{1,2}$. Then $\mathbb{E}[|u_i|] = \mu_1 \|X_i\|_2$ and $\mathbf{Var}[|u_i|] = O(\mathbb{E}^2[|u_i|])$, while $\mathbb{E}[||u||_1] = \mu_1 \sum_i \|X_i\|_2$. Thus $|u_i| = \Omega(\|X_i\|_2)$ and $\|u\|_1 = O(L_{1,2}(X))$ with large probability. If we run a heavy-hitters algorithm for $L_1(v)$ (e.g., [12]) we will find the heavy rows u_i. With minor modifications, one can achieve $T \subseteq HH_{\phi-\epsilon}(X)$ The space is $\text{poly}(\phi^{-1}\varepsilon^{-1}\log(nd))$.

Estimating $L_{k,\infty}$ and $L_{\infty,p}$: The tight space lower bounds follow via reductions from multi-party set-disjointness and L_∞. For $L_{\infty,p}$, this bound is achieved trivially (run L_p on each row). For $L_{k,\infty}$ where $k \geq 2$, observe that if there is a row X_i with $(L_\infty(X_i))^k \geq \varepsilon F_k(L_\infty(X))$, then $(L_\infty(X_i))^2 \geq \varepsilon^{2/k}(F_k \circ L_\infty)(X)^{2/k} \geq \varepsilon^{2/k} F_2(L_\infty(X))/n^{1-2/k}$, by Hölder's inequality. It follows that if we use *CountSketch* [11] to find all the $\varepsilon^{2/k}/n^{1-2/k}$-heavy hitters w.r.t. to F_2, we will find the coordinate in X_i realizing $L_\infty(X_i)$. It follows by subsampling entire rows at a time and running *CountSketch* on the substreams, we can estimate the sizes of all $S_t = \{i \mid (1+\varepsilon')^t \leq L_\infty(X_i) < (1+\varepsilon')^{t+1}\}$ for which $|S_t|(1+\varepsilon')^{kt} > L_{k,\infty}(X)\text{poly}(\varepsilon/\log n)$, and then use the analogous estimator as that for L_k in [19] to estimate $L_{k,\infty}$. Using [12], similar techniques work for $L_{1,\infty}$.

Estimating $L_{0,p}$: For any matrix X, $L_{0,p}(X) = L_{0,0}(X)$. Wlog assume that X is square. Each entry of X is bounded by $u = \text{poly}(n) \geq n$. Let q be a prime with $10u^2 \leq q < 20u^2$ and view X as a matrix over $GF(q)$. Let V be the $q \times n$ Vandermonde matrix over $GF(q)$, where row i is $(1, i, i^2, \ldots, i^{n-1}) \mod q$. Then any n rows of V are linearly independent. So for any i, if B_i is non-zero then at most $n-1$ rows v of V satisfy $\langle v, B_i\rangle = 0 \mod q$. Let v be a randomly chosen row of V. Then with probability at least $9/10$, for all i for which $B_i \neq 0$, $\langle v, B_i\rangle \neq 0 \mod q$. In this case we say that v is *good*. Let Alg be a $\text{poly}(\varepsilon^{-1}\log n)$-space algorithm [21] which outputs a $(1\pm\varepsilon)$-approximation to $L_0(u)$ of an input vector u with probability at least $5/6$. Given an update (i, j, x), we feed the pair $(i, v_j \cdot x)$ to Alg and return its output. This is a $(1\pm\varepsilon)$-approximation if Alg succeeds and v is good. The space is $\text{poly}(\varepsilon^{-1}\log n)$.

References

[1] "Open questions in data streams and related topics," in *IITK Workshop on Algorithms for Data Streams*, A. McGregor, Ed., 2006.

[2] N. Alon, Y. Matias, and M. Szegedy, "The space complexity of approximating the frequency moments," *JCSS*, vol. 58, no. 1, pp. 137–147, 1999.

[3] A. Andoni, K. DoBa, P. Indyk, and D. Woodruff, "Efficient sketches for earth-mover distance, with applications," in *FOCS*, 2009.

[4] A. Andoni, P. Indyk, and R. Krauthgamer, "Overcoming the ℓ_1 non-embeddability barrier: Algorithms for product metrics," in *SODA*, 2009.

[5] A. Andoni, T. S. Jayram, and M. Patrascu, "Lower bounds for edit distance and product metrics via poincaré-type inequalities," 2009.

[6] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar, "An information statistics approach to data stream and communication complexity," *JCSS*, vol. 68, no. 4, pp. 702–732, 2004.

[7] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan, "Counting distinct elements in a data stream," in *RANDOM*, 2002.

[8] L. Bhuvanagiri, S. Ganguly, D. Kesh, and C. Saha, "Simpler algorithm for estimating frequency moments of data streams," in *SODA*, 2006, pp. 708–713.

[9] A. Chakrabarti, Y. Shi, A. Wirth, and A. C.-C. Yao, "Informational complexity and the direct sum problem for simultaneous message complexity," in *FOCS*, 2001.

[10] A. Chakrabarti, S. Khot, and X. Sun, "Near-optimal lower bounds on the multi-party communication complexity of set disjointness," in *CCC*, 2003.

[11] M. Charikar, K. Chen, and M. Farach-Colton, "Finding frequent items in data streams," *TCS*, vol. 312, no. 1, pp. 3–15, 2004.

[12] G. Cormode and S. Muthukrishnan, "An improved data stream summary: the count-min sketch and its applications," *J. Algorithms*, vol. 55, no. 1, pp. 58–75, 2005.

[13] ——, "Space efficient mining of multigraph streams," in *PODS*. ACM, 2005, pp. 271–282.

[14] D. Feldman, M. Monemizadeh, C. Sohler, and D. Woodruff, "Coresets and sketches for high dimensional subspace approximation problems," 2009.

[15] P. Flajolet and G. N. Martin, "Probabilistic counting algorithms for data base applications," *JCSS*, vol. 31, no. 2, pp. 182–209, 1985.

[16] S. Ganguly, M. Bansal, and S. Dube, "Estimating hybrid frequency moments of data streams," in *FAW*, 2008.

[17] A. Gronemeier, "Asymptotically optimal lower bounds on the nih-multi-party information complexity of the and-function and disjointness," in *STACS*, 2009.

[18] P. Indyk, "Stable distributions, pseudorandom generators, embeddings, and data stream computation." *J. ACM*, vol. 53, no. 3, pp. 307–323, 2006.

[19] P. Indyk and D. P. Woodruff, "Optimal approximations of the frequency moments of data streams." in *STOC*, 2005.

[20] T. S. Jayram, "Hellinger strikes back: A note on the multi-party information complexity of AND," in *RANDOM*, 2009.

[21] D. Kane, J. Nelson, and D. Woodruff, "Revisiting norm estimation in data streams," 2009.

[22] J. Matousek, *Lectures on Discrete Geometry*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2002.

[23] M. Monemizadeh and D. Woodruff, "1-pass relative-error l_p sampling with applications," 2009.

[24] S. Muthukrishnan, "Data streams: Algorithms and applications," *Foundations and Trends in Theoretical Computer Science*, vol. 1, no. 2, 2005.

[25] N. Nisan, "Pseudorandom generators for space-bounded computation," *Combinatorica*, vol. 12, no. 4, pp. 449–461, 1992.

[26] D. Woodruff, "Efficient and private distance approximation in the communication and streaming models," Ph.D. dissertation, MIT, 2008.

[27] D. P. Woodruff, "Optimal space lower bounds for all frequency moments." in *SODA*, 2004, pp. 167–175.