
Schatten Norms in Matrix Streams: Hello Sparsity, Goodbye Dimension

Vladimir Braverman¹ Robert Krauthgamer² Aditya Krishnan¹ Roi Sinoff²

Abstract

Spectral functions of large matrices contain important structural information about the underlying data, and are thus becoming increasingly important. Many times, large matrices representing real-world data are *sparse* or *doubly sparse* (i.e., sparse in both rows and columns), and are accessed as a *stream* of updates, typically organized in *row-order*. In this setting, where space (memory) is the limiting resource, all known algorithms require space that is polynomial in the dimension of the matrix, even for sparse matrices. We address this challenge by providing the first algorithm whose space requirement is *independent of the matrix dimension*, assuming the matrix is doubly-sparse and presented in row-order. Our algorithms approximate the Schatten p -norms, which we use in turn to approximate other spectral functions, such as logarithm of the determinant, trace of matrix inverse, and Estrada index. We validate these theoretical performance bounds by numerical experiments on real-world matrices representing social networks. We further prove that multiple passes are unavoidable in this setting, and show extensions of our primary technique, including a trade-off between space requirements and number of passes.

1. Introduction

Large matrices are often used to represent real-world data sets like text documents, images and social networks, however analyzing them is increasingly challenging, as their

Authors are listed in alphabetical order. Full version : (Braverman et al., 2019) ¹Department of Computer Science, Johns Hopkins University, USA. ²Department of Computer Science & Applied Mathematics, The Weizmann Institute of Science, Israel. Correspondence to: Vladimir Braverman <vova@cs.jhu.edu>, Robert Krauthgamer <robert.krauthgamer@weizmann.ac.il>, Aditya Krishnan <akrish23@jhu.edu>, Roi Sinoff <roi.sinoff@gmail.com>.

Proceedings of the 37th International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

sheer size renders many algorithms impractical. Fortunately, in several application domains, input matrices are often very sparse, meaning that only a small fraction of their entries are non-zero. In fact, in applications related to natural language processing, image recognition, medical imaging and computer vision (e.g. (Ganitkevitch et al., 2013; Goyette et al., 2012; Liu et al., 2015)), the matrices are often doubly sparse, i.e., sparse in both rows and columns. Throughout, we say that a matrix is k -sparse if every row and every column has at most k non-zero entries. The current work devises new algorithms to analyze the *spectrum* (singular values) of such sparse matrices, aiming to achieve efficiency (storage requirement in streaming model) that depends on *matrix sparsity* k instead of *matrix dimension* n .

We focus on fundamental functions of the spectrum, called Schatten norms. Formally, the Schatten p -norm of a matrix $A \in \mathbb{R}^{m \times n}$, $m \geq n$ with singular values $\sigma_1 \geq \dots \geq \sigma_n \geq 0$ is defined for every $p \geq 1$ as

$$\|A\|_{S_p} := \left(\sum_{i=1}^n \sigma_i^p \right)^{1/p}.$$

This definition extends also to $p = 0, \infty$ by taking the limit. Frequently used cases include $p = 0$, representing the rank of A , and $p = 1, 2, \infty$, commonly known as the trace/nuclear norm, Frobenius norm, and spectral/operator norm, respectively. Schatten norms are often used as surrogates for the spectrum, as explained in (Zhang et al., 2015; Kong and Valiant, 2016; Napoli et al., 2016; Khetan and Oh, 2019), and some specific cases have applications in optimization, image processing, and differential privacy etc. (Xie et al., 2016; Musco et al., 2018). For simplicity, we focus on the case $m = n$.

For a positive semidefinite (PSD) matrix A , the Schatten norms can be easily used to approximate other important spectral functions. One example is the trace of matrix inverse $\text{Tr}(A^{-1})$, which is used for image restoration, for counting triangles in graphs, to measure the uncertainty in data collections, and to bound the total variance of unbiased estimators (see e.g. (Wu et al., 2016; Chen, 2016; Han et al., 2017; Braverman, 2019)). Another example is the Estrada index, which has applications in chemistry, physics, network theory and information theory (see the

survey (Gutman et al., 2011)). A third example is the logarithm of the determinant $\log \det(A)$, used in many machine learning tasks, such as Bayesian learning, kernel learning, Gaussian processes, tree mixture models, spatial statistics and Markov field models (see e.g. (Han et al., 2015; Ubaru et al., 2017; Ubaru and Saad, 2017; Han et al., 2017)). Thus, our results for Schatten norm have further applications.

As matrices in many real-world applications are often very large, storing these entire matrices in working memory can be impractical, and thus analyzing them has become increasingly challenging. As a result, the data-stream model has emerged as a convenient model for how these data-sets are accessed in practice. In this model, the input matrix $A \in \mathbb{R}^{m \times n}$ is presented as a sequence of items/updates. In one common setting, the *turnstile* model, each update has the form (i, j, δ) for $\delta \in \mathbb{Z}$ and represents an update $A_{ij} \leftarrow A_{ij} + \delta$. In another common setting, the *row-order* model, items (i, j, A_{ij}) arrive in lexicographic order of their location (i, j) , providing directly the entry $A_{ij} \in \mathbb{Z}$ in that location. In both models, unspecified entries are 0 by convention, which is very effective for sparse matrices.

Row-order is a common access pattern for external memory algorithms. When the data is too large to fit into working memory and has to be “streamed” into memory in some pattern, it is useful to assume that algorithms can make multiple, albeit few, passes over the input data. For a thorough study of external memory algorithms, including motivation for the row-order model and for multiple passes, see (Gibbons and Matias, 1999; Vitter, 2001; Liberty, 2013).

Designing small-space algorithms for estimating Schatten norms of an input matrix in the data-stream model was investigated recently for various matrix classes and stream types (Clarkson and Woodruff, 2009; Andoni and Nguyen, 2013; Li et al., 2014; Li and Woodruff, 2016a;b; 2017; Braverman et al., 2018). However, all known algorithms require space that is polynomial in n , the matrix dimension, even if the matrix is highly sparse and the stream type is favorable, say row-order. A natural question then is:

Q: Does any streaming model admit algorithms for computing Schatten norms of a matrix, using storage requirement independent of the matrix dimension?

We answer this question in the affirmative for k -sparse matrices presented in row-order and all even integers p . Our algorithms extend to all integers $p \geq 1$ if the input matrix is PSD.

1.1. Main Results

Upper and Lower Bounds for Row-Order Streams. Our main result is a new algorithm for approximating the

Schatten p -norm (for even p) of a k -sparse matrix streamed in row-order, using $O(p)$ passes and $\text{poly}(k^p/\varepsilon)$ space (independent of the matrix dimension). We exploit the matrix sparsity in a series of novel algebraic lemmas to reduce the exponent in the space bound and achieve the following theorem, whose proof appears in Section 4 in the full version.

Theorem 1.1. *There exists an algorithm that, given $p \in 2\mathbb{Z}_{\geq 2}$, $\varepsilon > 0$ and a k -sparse matrix $A \in \mathbb{R}^{n \times n}$ streamed in row-order, makes $\lfloor p/4 \rfloor + 1$ passes over the stream using $O_p(\varepsilon^{-2} k^{3p/2-3})$ words of space, and outputs $\tilde{Y}(A)$ that $(1 \pm \varepsilon)$ -approximates $\|A\|_{S_p}^p$ with probability at least $2/3$.*

Here and throughout, we write $\tilde{O}(f)$ as a shorthand for $O(f \cdot \log^{O(1)} n)$ where n is the dimension of the matrix, and write $O_d(f)$ when the hidden constant might depend on the parameter d . We assume the matrix entries are integers bounded by $\text{poly}(n)$, and thus often count space in words, each having $O(\log n)$ bits. We denote by $\lceil p \rceil_4$ the smallest multiple of 4 that is greater than or equal to p , and similarly by $\lfloor p \rfloor_4$ the largest multiple of 4 that is smaller than or equal to p .

Theorem 1.1 provides a multi-pass algorithm whose space complexity depends only on the sparsity of the input matrix. A natural question is whether one can achieve a similar dependence also for one-pass algorithms, but our next theorem (proved in Section 6 in the full version) shows that such algorithms require $\text{poly}(n)$ bits of space, even for $O(1)$ -sparse matrices. Thus, multiple passes are necessary to achieve storage requirement that is independent of the matrix dimensions, even for sparse matrices.

Theorem 1.2. *For every $p \in 2\mathbb{Z}_{\geq 2}$ there is $\varepsilon(p) > 0$ such that every algorithm that makes one pass over an $O_p(1)$ -sparse matrix $A \in \mathbb{R}^{n \times n}$ streamed in row-order, and then $(1 \pm \varepsilon(p))$ -approximates $\|A\|_{S_p}^p$ with probability at least $2/3$, must use $\Omega(n^{1-4/\lceil p \rceil_4})$ bits of space.*

We further extend our primary algorithmic technique (from Theorem 1.1) in several different ways, and obtain improved algorithms for special families of matrices, algorithms in the more general turnstile model, and algorithms with a trade-off between the number of passes and the space requirement, as explained later in this section. Table 1 summarizes our bounds for row-order streams, and compares them to those derived from previous work (when applicable).

Applications for Approximating Schatten Norms. We show in Section 8 two settings where, under certain simplifying conditions, our algorithms can be used to approximate other functions of the spectrum, and even weakly recover the entire spectrum. The basic idea is that it suffices to compute only a few Schatten norms, in which case our algorithms for k -sparse matrices in row-order streams can

Table 1: Space bounds for Schatten norms (for even p) of k -sparse matrices in row-order streams. Upper bounds are counted in words. Lower bounds are in bits, for a suitable $\varepsilon(p) > 0$.

Which p	Space Bound	Ref.	Comments
$p > 4$	$\tilde{O}_{p,\varepsilon}(k^{O(p)}n^{1-4/\lceil p \rceil 4})$	(Braverman et al., 2018)	one-pass
	$O_{p,\varepsilon}(k^{3p/2-3})$	Thm. 1.1	$\lfloor p/4 \rfloor + 1$ passes
	$O_{p,\varepsilon}(k^{2ps}n^{1-1/s})$	Thm. 5.3 in full version	$\lfloor \frac{p}{2(s+1)} \rfloor + 1$ passes
	$\Omega(n^{1-4/\lceil p \rceil 4})$	Thm. 1.2	one-pass, $k = O_p(1)$
	$\Omega_t(k^{p/2-2})$	(Braverman et al., 2018)	t passes, $k \leq n^{2/p}$
$p = 4$	$\tilde{O}_{p,\varepsilon}(k)$	(Braverman et al., 2018)	one-pass
	$O_p(\varepsilon^{-2})$	Thm. 7.2 in full version	one-pass, for all $k \leq n$

be used, and then the overall algorithm requires only small space (depending on k).

The first setting considers spectral sums (i.e., $\sum_i f(\sigma_i)$ for some f) of PSD matrices. We use an idea from (Boutsidis et al., 2017) to show that for a PSD input $A \in \mathbb{R}^{n \times n}$ whose eigenvalues lie in an interval $[\theta, 1]$, one can $(1 \pm 2\varepsilon)$ -approximate $\log \det(A)$ and $\text{Tr}(A^{-1})$ using the first $\lceil \frac{1}{\theta} \log(\frac{1}{\varepsilon}) \rceil$ (integer) Schatten norms. We further show that given a Laplacian matrix whose eigenvalues lie in an interval $[0, \theta]$, one can $(1 \pm 2\varepsilon)$ -approximate the Estrada index using the first $\lceil (e\theta + 1) \log \frac{1}{\varepsilon} \rceil$ (integer) Schatten norms.

The second setting considers recovering the spectrum of PSD matrices. We use an idea from (Kong and Valiant, 2016) to approximate the spectrum of a PSD matrix with eigenvalues in the interval $[0, 1]$ to within L_1 -distance εn , using the first $O(1/\varepsilon)$ Schatten norms.

Experiments. We validated our row-order algorithm on real-world matrices representing academic collaboration network graphs. The experiments show that the space needed to approximate the Schatten 6-norm of these matrices is much smaller than the theoretical bound, and that the algorithm is efficient also for larger values of p . In fact, the matrices in our experiments obey the sparsity requirement in every row, but their columns are sparse only on average. Finally, we also experimented whether the algorithm is robust to noise, and found that it is indeed effective also for nearly-sparse matrices. Finally, our experiments validate that the storage requirement is independent of the matrix dimensions. See Section 9 for details.

1.2. Extensions of Main Results

We give a number of extensions to our main result, all of the details of which we defer to the full version.

Extension I: Fewer Passes. We show how to generalize our algorithmic technique to use fewer passes over

the stream, albeit requiring more space. Our method uses a novel variant of importance sampling which we refer to as *set-sampling*, and attains the following pass-space trade-off. For any integer $s \geq 2$, our algorithm makes $t(s) = \lfloor \frac{p}{2(s+1)} \rfloor + 1$ passes over the stream using $O_p(\varepsilon^{-3}k^{2ps}n^{1-1/s})$ words of space, and outputs a $(1 \pm \varepsilon)$ -approximation to $\|A\|_{S_p}^p$ for $p \in 2\mathbb{Z}_{\geq 2}$.

Extension II: Turnstile Streams. We design an algorithm for *turnstile* streams with an additional $\tilde{O}(\varepsilon^{-O(p)}k^{3p/2-3}n^{1-2/p})$ factor in their space complexity compared to our algorithm for row-order streams. The term $O(n^{1-2/p})$ is quite expected here, since the space complexity for estimating ℓ_p -norms of vectors in turnstile streams using t passes is $\Omega(\frac{n^{1-2/p}}{t})$, and our turnstile algorithm makes $p + 1$ passes over the stream.

Extension III: Special Matrix Families. We give improved bounds for special families of k -sparse matrices that may be of potential interest. For Laplacians of undirected graphs with degree at most $k \in \mathbb{N}$, we show $(1 \pm \varepsilon)$ -approximation of the Schatten p -norm using space $O_p(\varepsilon^{-2}k^{p/2-1})$ in $p/2$ passes over a row-order stream. Additionally, for matrices whose non-zero entries lie in an interval $[\alpha, \beta]$ for $\alpha, \beta \in \mathbb{R}^+$, we can get nearly-tight upper bounds – our algorithm uses space $O_p(\varepsilon^{-2}k^{p/2-1}(\beta/\alpha)^{p/2-2})$, which for $\alpha = \beta = 1$ (i.e., 0 – 1 entries) is nearly tight with the $\Omega(k^{p/2-2})$ multi-pass lower bound of (Braverman et al., 2018).

Schatten 4-norm. We show a simple one-pass algorithm for $(1 \pm \varepsilon)$ -approximating the Schatten 4-norm of *any* matrix (not necessarily sparse) given in a row-order stream, using only $\tilde{O}_p(\varepsilon^{-2})$ words of space. This improves a previous $\tilde{O}_p(\varepsilon^{-2}k)$ bound from (Braverman et al., 2018).

1.3. Technical Overview

Our algorithms are based on the principle of importance sampling. The basic version outlined below easily achieves storage requirement $\text{poly}(k^p/\varepsilon)$ (independent of the matrix dimension), and our full analysis, which specifies the hidden constants in the exponents, including some non-trivial optimizations. When describing our algorithms, it will be convenient to first design an unbiased estimator for $\|A\|_{S_p}^p$, and then implement this estimator by a streaming algorithm with few passes. To the best of our knowledge, these are the first streaming algorithms for Schatten norms that use adaptive sampling, i.e., the sampling probabilities in every pass depend on observations at the preceding pass.

For an integer $p \in 2\mathbb{Z}_{\geq 1}$ and $q := p/2$, the Schatten p -norm of a matrix $A \in \mathbb{R}^{n \times n}$ can be expressed as

$$\|A\|_{S_p}^p = \text{Tr}((AA^\top)^q) \quad (1.1)$$

$$= \sum_{i_1, \dots, i_q \in [n]} \langle a_{i_1}, a_{i_2} \rangle \langle a_{i_2}, a_{i_3} \rangle \dots \langle a_{i_q}, a_{i_1} \rangle \quad (1.2)$$

where a_i is the i^{th} row of matrix A . We can interpret (1.2) as a sum over cycles of q inner-products between rows of A . We refer to these informally as *cycles*, and *assign* each cycle to one of the rows participating in that cycle. Hence, we can write the Schatten p -norm as a sum $\sum_{i=1}^n z_i$, where z_i is the cumulative weight of all cycles assigned to row i .

Our estimator starts by sampling a row $i \in [n]$ with probability proportional to the *heaviest* cycle assigned to row i (i.e., largest contribution to z_i). It then sample one cycle assigned to i with probability proportional to the weight of the cycle (actually performed by an iterative process with $q/2 = p/4$ stages). Since the rows *and* columns are sparse, each row cannot participate in too many cycles (because it is orthogonal to every row with a disjoint support), and thus the number of cycles assigned to each row i is roughly $k^{O(p)}$. It follows that sampling the first row with probability proportional to the heaviest contributing cycle is a good approximation (within factor $k^{O(p)}$) to sampling proportionally to z_i , the actual contribution of row i to $\sum_{i \in [n]} z_i = \|A\|_{S_p}^p$.

Sampling a row with probability proportional to its heaviest contributing cycle depends on the assigning process. A natural assignment is to assign every cycle to the row with largest l_2 -norm participating in that cycle (breaking ties arbitrarily), because then by the Cauchy-Schwarz inequality, the heaviest contributing cycle to row i is simply $\|a_i\|_2^p$.

The above algorithm is based on two technical ideas. The first one is to use importance sampling to sample the first (“seed”) row, which already suffices to achieve a dimension-independent space bound $\text{poly}(k^p/\varepsilon)$. Indeed, a streaming algorithm with $O(p)$ passes can easily compute the contribution z_i of this row (the total weight of all

cycles that contain this row). The second idea is to use importance sampling repeatedly (after picking a seed row) in order to sample a single cycle that contains the seed row. Indeed, a streaming algorithm with $O(p)$ passes can iteratively “grow” (randomly) a single cycle around the seed row. Our analysis bounds the variance of this estimator by $O(k^{3p/2-4})$, which gives our desired result.

In the row-order model, this estimator can be implemented easily using weighted reservoir sampling (Vitter, 1985; Braverman et al., 2015), as discussed in Section 4. However, implementing it in turnstile streams is more challenging. Using approximate L_p -samplers presented in (Monemizadeh and Woodruff, 2010), we build an approximate cascaded $L_{p,2}$ -norm sampler,¹ to sample each row i with probability proportional to $\|a_i\|_2^p$. Additionally, we use the Count-Sketch data structure to recover rows and sample cycles once we have sampled the “seed” row. This allows us to implement the estimator in turnstile data streams with an additional $\tilde{O}(\varepsilon^{-O(p)} n^{1-2/p})$ factor in the space complexity attributed to the cascaded norm sampler, and an additional $O_p(k^{3p/2-3})$ factor that comes from approximating the sampling probabilities (compared to row-order in which the sampling probabilities can be recovered exactly).

In Section 5 we generalize the design of the importance sampling estimator. Instead of assigning every cycle to a single row that participates in it, every cycle is mapped to s rows participating in it, for a parameter $s \in \mathbb{N}$. These s rows are used to split the cycle into roughly s segments, such that in each segment, the heaviest row (by l_2 norm) is one of the s assigned rows. The algorithm samples s “seed” rows and then computes the total weight of all the cycles assigned to these s rows (or alternatively samples one such cycle). The length of these segment is shown to decrease linearly with s , hence they can be computed with fewer passes. However, the algorithm needs to sample more indices in order to ensure that each cycle has a sufficiently large probability of being “hit” s times. This tension leads to a trade-off between passes and space.

1.4. Previous and Related Work

The bilinear sketching algorithm in (Li et al., 2014) was the first non-trivial algorithm for Schatten p -norm estimation in turnstile streams. It requires only one-pass over the data and uses $O(\varepsilon^{-2} n^{2-4/p})$ words of space.² Their algorithm uses $O(\varepsilon^{-2})$ independent $G_1 A G_2^\top$ sketches, where $G_1, G_2 \in \mathbb{R}^{t \times n}$ are matrices with i.i.d. Gaussian entries and $t = O(n^{1-2/p})$.

¹The cascaded $L_{p,2}$ -norm of a matrix $A \in \mathbb{R}^{n \times m}$ for $p \geq 0$ is $(\sum_{i=1}^n \|a_i\|_2^p)^{1/p}$.

²They also showed an $\Omega(n^{2-4/p})$ lower bound for the dimension of bilinear sketching for approximating $\|A\|_{S_p}^p$ for all $p \geq 2$.

Inspired by this sketch, (Braverman et al., 2018) gave an almost quadratic improvement in the space complexity if the algorithm is allowed to make multiple passes over the data. Their estimator uses matrices $G_2, \dots, G_p \in \mathbb{R}^{t \times n}$ with i.i.d. Gaussian entries and Gaussian vector $g_1 \in \mathbb{R}^n$ to output $g_1^\top A G_2^\top G_2 A \dots G_p A g_1$. This estimate can be constructed in $p/2$ passes of the data and requires $O(\varepsilon^{-2})$ independent copies each using only $t = O(n^{1-\frac{1}{p-1}})$ space.

Restricting the input matrix to be $O(1)$ -sparse allows for quadratic improvement in the space complexity of one-pass algorithms as shown in (Li and Woodruff, 2016a). They show that sampling $O(n^{1-2/p})$ rows and storing them approximately using small space (since each row is sparse) is sufficient to $(1 + \varepsilon)$ -approximate the Schatten p -norm by exploiting the fact that rows cannot “interact” with one another “too much” because of the sparsity restriction.

If we restrict the data stream to be row-order, then we can reduce the dependence on p in all the above algorithms by a factor of 2. As noted in (Braverman et al., 2018), since $A^\top A = \sum_i a_i a_i^\top$ (where a_i is the i th row of A) one can apply the above algorithms to $A^\top A$ instead of A by updating it with the outer product of every row with itself. Since $\|A^\top A\|_{S_{p/2}}^{p/2} = \|A\|_{S_p}^p$ (for even p values), the output is as desired and the dependence on p reduces by a factor of 2.

Lower Bounds. Every t -pass algorithm designed for turnstile streams requires $\Omega(n^{1-2/p}/t)$ bits, which follows by injecting the F_p -moment problem (see (Gronemeier, 2009; Jayram, 2009)) into the diagonal elements. Li and Woodruff (Li and Woodruff, 2016a) showed that restricting the algorithm to a single pass over the turnstile stream, leads to a lower bound $\Omega(n^{1-\varepsilon})$ bits for every fixed $\varepsilon > 0$ and $p \notin 2\mathbb{Z}_{\geq 2}$, even if the input matrix is $O(1)$ -sparse.³ Later (Braverman et al., 2018) proved that $\Omega(n^{1-\varepsilon})$ bits are required for $p \notin 2\mathbb{Z}_{\geq 2}$ even in row-order streams. In addition, they showed (Theorem 5.4 in Arxiv version) that t passes over row-order streams require space $\Omega(n^{1-4/p}/t)$ bits, however these matrices are actually $\Omega(n^{2/p})$ -sparse (and not $O(1)$ -sparse as may be understood from Table 2 therein). A simple adaptation of that result yields an $\Omega(k^{p/2-2}/t)$ space lower bound for k -sparse input matrices ($k \leq n^{2/p}$).

2. Preliminaries

Importance Sampling. Our main algorithmic technique is inspired by the importance sampling framework, as formulated by the following theorem, proved in the full version.

³ They also showed that for $p \in 2\mathbb{Z}_{\geq 2}$, single-pass algorithms require $\Omega(n^{1-2/p})$ bits even if all non-zeros in the input matrix are constants.

Theorem 2.1 (Importance Sampling). *Let $z = \sum_{i \in [n]} z_i \geq 0$ be a sum of n reals. Let the random variable \hat{Z} be an estimator computed by sampling a single index $i \in [n]$ according to the probability distribution given by $\{\tau_i\}_{i=1}^n$ and setting $\hat{Z} \leftarrow \frac{z_i}{\tau_i}$. If for some parameter $\lambda \geq 1$, each $\tau_i \geq \frac{|z_i|}{\lambda z}$, then $\mathbb{E}[\hat{Z}] = z$ and $\text{Var}(\hat{Z}) \leq (\lambda z)^2$.*

3. An Estimator for Schatten p -Norm for $p \in 2\mathbb{Z}_{\geq 2}$

This section introduces our importance sampling estimator for Schatten p -norms.

3.1. Preliminaries

Fix a matrix $A \in \mathbb{R}^{n \times n}$ and $p \in 2\mathbb{Z}_{\geq 2}$. For a row a_i , we define the set of its *neighboring rows* $N(i) := \{l \in [n] : \text{supp}(a_i) \cap \text{supp}(a_l) \neq \emptyset\}$. In addition, we denote the set of neighboring rows of a_j that have smaller length than row a_i by $N_S^i(j) := \{l \in N(j) : \|a_l\|_2 \leq \|a_i\|_2\}$.

Building on this, we introduce notation for certain “paths” of rows. Fixing some row indices $i, i_1 \in [n]$ and an integer $t \geq 2$, we then define $\Gamma_S^i(i_1, t) := \{(i_1, \dots, i_t) : i_2 \in N_S^i(i_1), \dots, i_t \in N_S^i(i_{t-1})\}$.

We further define the weights of “paths” of inner products: given an integer $t \geq 2$ and indices $i_1, \dots, i_t \in [n]$, let $\sigma(i_1, \dots, i_t) := \langle a_{i_1}, a_{i_2} \rangle \langle a_{i_2}, a_{i_3} \rangle \dots \langle a_{i_{t-1}}, a_{i_t} \rangle$.

Recall from (1.2) that the Schatten p -norm of $A \in \mathbb{R}^{n \times n}$ can be expressed in terms of the product of inner products of the rows of A . Using the above notation we manipulate it as follows.

$$\begin{aligned} \|A\|_{S_p}^p &= \text{Tr}((AA^\top)^q) = \sum_{i_1, \dots, i_q \in [n]} \sigma(i_1, \dots, i_q, i_1) \\ &= \sum_{i_1} \sum_{\substack{(i_1, \dots, i_{q-1}) \\ \in \Gamma_S^{i_1}(i_1, q-1)}} \sum_{i_q \in N_S^{i_1}(i_1)} c(i_1, \dots, i_q) \sigma(i_1, \dots, i_q, i_1) \end{aligned} \quad (3.1)$$

where $1 \leq c(i_1, \dots, i_q) \leq q$ is the number of times the sequence (i_1, \dots, i_q, i_1) or a cyclic shift of the sequence appears in the first equation.

3.2. The Estimator

Our estimator is an importance sampling estimator for the quantity in (3.1). To define it, we need the following quan-

ities:

$$\mathcal{S} := \bigcup_{i \in [n]} \Gamma_S^i(i, q-1)$$

$$\forall (i_1, \dots, i_{q-1}) \in \mathcal{S},$$

$$z_{(i_1, \dots, i_{q-1})} := \sum_{i_q \in N_S^{i_1}(i_1)} c(i_1, \dots, i_q) \sigma(i_1, \dots, i_q) \langle a_{i_q}, a_{i_1} \rangle$$

$$z := \sum_{(i_1, \dots, i_{q-1}) \in \mathcal{S}} z_{(i_1, \dots, i_{q-1})} = \|A\|_{S_p}^p$$

Our importance sampling estimator, for the sum z , samples quantities $z_{(i_1, \dots, i_{q-1})}$ indexed by $(i_1, \dots, i_{q-1}) \in \mathcal{S}$ in $q-1$ steps. In the first step, it samples row $i_1 \in [n]$ with probability $\frac{\|a_{i_1}\|_2^p}{\sum_j \|a_j\|_2^p}$. In each step $2 \leq t \leq q-1$, conditioned on sampling i_{t-1} in step $t-1$ it samples row $i_t \in N_S^{i_1}(i_{t-1})$ with probability $p_{i_{t-1}}^{i_1}(i_t) := \frac{|\langle a_{i_{t-1}}, a_{i_t} \rangle|}{\sum_{i \in N_S^{i_1}(i_{t-1})} |\langle a_{i_{t-1}}, a_i \rangle|}$.

Overall, a sequence $(i_1, \dots, i_{q-1}) \in \mathcal{S}$ is sampled with probability $\tau_{(i_1, \dots, i_{q-1})} = \frac{\|a_{i_1}\|_2^p}{\sum_j \|a_j\|_2^p} \prod_{t=2}^{q-1} p_{i_{t-1}}^{i_1}(i_t)$, and the output estimator is $Y(A) := \frac{1}{\tau_{(i_1, \dots, i_{q-1})}} \cdot z_{(i_1, \dots, i_{q-1})}$.

We prove in the below theorem (using new technical lemmas that we call “projection lemmas”), that the importance sampling estimator $Y(A)$ is an unbiased estimator with a small variance. We provide an improved version of the theorem below for two special families of k -sparse $n \times n$ matrices: (i) Laplacians of undirected graphs, denoted by \mathcal{L}_n and (ii) matrices whose non-zero entries lie in an interval $[\alpha, \beta]$ for parameters $0 < \alpha \leq \beta$ which we denote by $\mathcal{C}_{\alpha, \beta}^{n \times n}$.

Theorem 3.1. *For every $p \in 2\mathbb{Z}_{\geq 2}$ and a k -sparse matrix $A \in \mathbb{R}^{n \times n}$, the estimator $Y(A)$ given in Section 3.2 satisfies $\mathbb{E}[Y(A)] = \|A\|_{S_p}^p$ and $\text{Var}(Y(A)) \leq O_p(k^{\frac{3p}{2}-4}) \|A\|_{S_p}^{2p}$.*

4. Implementing the Estimator: Row-Order and Turnstile Streams

In this section we show how to implement the importance sampling estimator defined in Section 3.2 in two different streaming models, row-order and turnstile streams. We start by stating two theorems that bound the space complexity of implementing the estimator in row-order streams. The first one is our main result from the Introduction, and applies to all k -sparse matrices. The second theorem considers special families of k -sparse matrices.

Theorem 1.1. *There exists an algorithm that, given $p \in 2\mathbb{Z}_{\geq 2}$, $\varepsilon > 0$ and a k -sparse matrix $A \in \mathbb{R}^{n \times n}$ streamed in row-order, makes $\lfloor p/4 \rfloor + 1$ passes over the stream using $O_p(\varepsilon^{-2} k^{3p/2-3})$ words of space, and outputs $\bar{Y}(A)$ that $(1 \pm \varepsilon)$ -approximates $\|A\|_{S_p}^p$ with probability at least $2/3$.*

Theorem 4.1. *There exists an algorithm that, given $p \in 2\mathbb{Z}_{\geq 2}$, $\varepsilon > 0$, and a k -sparse matrix $A \in \mathcal{L}_n$ streamed in row-order, makes $\lfloor p/4 \rfloor + 1$ passes over the stream using $O_p(\varepsilon^{-2} k^{p/2})$ words of space, and outputs $\bar{Y}(A)$ that $(1 \pm \varepsilon)$ -approximates $\|A\|_{S_p}^p$ with probability at least $2/3$. If instead the k -sparse matrix A is from $\mathcal{C}_{\alpha, \beta}^{n \times n}$ for $0 < \alpha \leq \beta$, then the space bound is $O_p(\varepsilon^{-2} k^{p/2-1} (\beta/\alpha)^{p/2-2})$ words.*

We also show that the estimator defined in Section 3.2 can be implemented in turnstile streams in $p/2 + 3$ passes over the stream.

Theorem 4.2. *There exists an algorithm that, given $p \in 2\mathbb{Z}_{\geq 2}$, $\varepsilon > 0$ and a k -sparse matrix $A \in \mathbb{R}^{n \times n}$ streamed in a turnstile fashion, makes $p/2 + 3$ passes over the stream using $O_p(k^{3p-6} n^{1-\frac{2}{p}} (\varepsilon^{-1} \log n)^{O(p)})$ words of space, and outputs $\bar{Y}(A)$ that $(1 \pm \varepsilon)$ -approximates $\|A\|_{S_p}^p$ with probability at least $2/3$.*

Outline. At a high level, the algorithms in all three theorems are similar, and compute multiple copies of the estimator defined in Section 3.2 in parallel and output their average (to reduce the variance). The algorithms differ in the number of copies. The first stage samples and stores a “seed” row which we will denote by a_{i_1} . Each subsequent stage $1 < t < q$ stores two values: a row index i_t (and row a_{i_t} itself) and an interim estimate $Y_t := \sigma(i_1, \dots, i_t)$. The final stage q computes and outputs $\sum_{i_q \in N_S^{i_1}(i_1)} Y_{q-1} \cdot \langle a_{i_q}, a_{i_1} \rangle c(i_1, \dots, i_q)$, where $1 \leq c(i_1, \dots, i_q) \leq q$ is as defined in (3.1).

The estimator is relatively easy to implement in row-order streams using $\lfloor p/4 \rfloor + 1$ passes and $O_p(\varepsilon^{-2} k^{3p/2-3})$ words of space. In turnstile streams however, the estimator is more difficult to implement. The first technical roadblock is sampling the first, “seed” row $i_1 \in [n]$ with probability proportional to $\frac{\|a_{i_1}\|_2^p}{\sum_j \|a_j\|_2^p}$. We use approximate samplers for turnstile streams to get around this roadblock. Approximate samplers introduce a multiplicative (relative) error and an additive error in the sampling probability, which need to be accounted for when analyzing the algorithm that uses the sampler. The second technical roadblock is computing inner-products between sampled rows which we solve by using a Count-Sketch data-structure to recover the relevant information of the rows.

Using the two subroutines we can implement the estimator in Section 3.2 in $p + 1$ passes of the stream in space $O_p(k^{3p-6} n^{1-2/p} (\varepsilon^{-1} \log n)^{O(p)})$. The additional $\tilde{O}(n^{1-2/p})$ space complexity factor is introduced by the approximate $L_{p,2}$ -sampler.

5. Pass-Space Trade-off

In the full version, we show a trade-off in the number of passes and required space to compute the Schatten p -norm of k sparse matrices streamed in row-order.

6. Lower Bound for One-Pass Algorithms in the Row-Order Model

In the full version, we show a space lower bound of $\Omega(n^{1-4/\lfloor p \rfloor 4})$ bits for one-pass algorithms and even p values in the row-order model. This lower bound holds even if the matrix is promised to be $O(1)$ -sparse. Our main technical contribution is the analysis of even p values in a reduction presented in (Li and Woodruff, 2016a), based on the Boolean Hidden Hypermatching (Verbin and Yu, 2011; Bury and Schwiegelshohn, 2015). Our bound is closely related to the $\Omega(n^{1-1/\varepsilon})$ bits lower bound for $p \notin 2\mathbb{Z}$, proved in (Braverman et al., 2018), and is also nearly tight with the upper bound from the same paper.

7. $O_\varepsilon(1)$ -Space Algorithm for Schatten 4-Norm of General Matrices

In the full version, we present an $O(1/\varepsilon^2)$ -space algorithm for $(1 + \varepsilon)$ -approximation of the Schatten 4-norm in the row-order model.

8. Applications

In this section we present two applications of our Schatten-norm algorithms to compute other functions of the spectrum by approximating these functions using low-degree polynomials and spectrum recovery.

8.1. Approximating Spectral Sums of Positive Definite Matrices

We demonstrate how our Schatten-norm estimators can be used to approximate commonly used spectral functions of sparse matrices presented as a data stream. We consider three different spectral functions, log-determinant, trace of matrix inverse and Estrada index of a Laplacian matrix, that all belong to the class of spectral sums, as defined below. These results apply to sparse matrices that are either positive definite (PD), positive semidefinite (PSD).

Definition 8.1 (Spectral Sums (Han et al., 2017)). Given a function $f : \mathbb{R} \rightarrow \mathbb{R}$ and a matrix $A \in \mathbb{R}^{n \times n}$ with real eigenvalues $\lambda_1, \dots, \lambda_n$, a spectral sum is defined as $S_f(A) = \text{Tr}(f(A)) = \sum_{i=1}^n f(\lambda_i)$. When $f(x) = \log x$, the sum is known as log-determinant, when $f(x) = 1/x$ it is known as the trace of the matrix inverse, and when $f(x) = \exp(x)$ it is known as Estrada index.

Theorem 8.2. For every spectral function S_f from Table 2,

Table 2: Table listing results of Theorem 8.2.

S_f	I_f	m_f	W_f
Log-Det	$[\theta, 2)$	$\lceil \frac{1}{\theta} \cdot \log \frac{1}{\varepsilon} \rceil$	$\frac{1}{\varepsilon^2} k^{\frac{3m_f}{2}-3}$
Trace of Inv.	$[\theta, 2)$	$\lceil \frac{1}{\theta} \cdot \log \frac{1}{\varepsilon} \rceil$	$\frac{1}{\varepsilon^2} k^{\frac{3m_f}{2}-3}$
Estrada Index	$[0, \theta]^4$	$\lceil (e\theta + 1) \log \frac{1}{\varepsilon} \rceil$	$\frac{1}{\varepsilon^2} k^{m_f/2}$

there is an algorithm with the following guarantee. Given as input $\varepsilon, \theta > 0$, and a k -sparse matrix $A \in \mathbb{R}^{n \times n}$ presented as a row-order stream and whose eigenvalues all lie in the interval I_f given in the table, the algorithm makes $\lfloor m_f/4 \rfloor + 1$ passes over the stream using $O_{m_f}(W_f)$ words of space and outputs an estimate $\rho(A)$ such that $\Pr[\rho(A) \in (1 \pm 2\varepsilon)S_f(A)] \geq 2/3$.

At a high level, the proof follows that of (Boutsidis et al., 2017), who present a time-efficient algorithm for approximating the log-determinant of PD matrices. Our algorithm approximates each of the terms in the truncated Taylor expansion (of each function) *separately*, and thus we need all the Taylor expansion coefficients to be non-negative, which indeed applies for these three spectral functions.

8.2. Approximating the Spectrum of PSD matrices

We present an application of our algorithm to (weakly) estimate the spectrum of a matrix, with eigenvalues bounded in $[0, 1]$ using approximations of a “few” Schatten norms of the matrix. This is based on the work of Cohen-Steiner et al. (2018) on approximating the spectrum of a graph which is in turn based on insightful work by Kong and Valiant (2016) on approximately recovering a distribution from its moments using the Moment Inverse method.

Fix a PSD matrix $A \in \mathbb{R}^{n \times n}$ with eigenvalues $1 \geq \lambda_1 \geq \dots \geq \lambda_n$ and define the l -th moment of the spectrum to be $\frac{1}{n} \|A\|_{S_l}^l = \frac{1}{n} \sum_{i \in [n]} \lambda_i^l$. Cohen-Steiner et al. show that estimating $O(1/\varepsilon)$ moments of A up to multiplicative error $O(\varepsilon)$ is sufficient to estimate the spectrum of A within earth-mover distance $O(\varepsilon)$. It is well-known that the L_1 distance between two sorted vectors of length n is exactly n times the earth-mover distance between the corresponding point-mass distributions. Hence, for an error parameter $\varepsilon > 0$ and parameter $s = \frac{C}{\varepsilon}$ (where $C > 0$ is an absolute constant), given a k -sparse PSD matrix $A \in \mathbb{R}^{n \times n}$ that is streamed in row-order and whose eigenvalues are in the range $[0, 1]$, one can use our row-order algorithm and the recovery scheme of Cohen-Steiner et al. to recover the spectrum of A within L_1 distance $O(\varepsilon n)$ using space $O(k^{3s/2-3} \exp(-C'\varepsilon))$ for some absolute constant $C' > 0$ and using $\lfloor s/4 \rfloor + 1$ passes over the stream.

⁴If the underlying graph is unweighted then the largest eigenvalue is bounded by the degree, i.e. $\theta \leq 2k$.

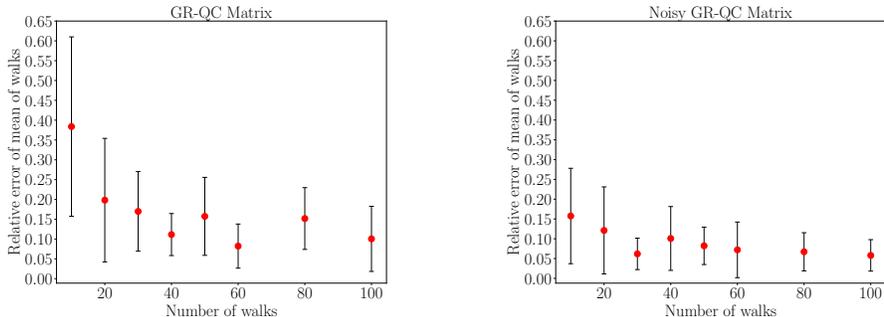


Figure 1: Relative error of Row-Order Algorithm for Schatten 6-norm of arXiv General Relativity and Quantum Cosmology Collaboration Network: Vary number of walks and plot relative error of the mean of the walks.

9. Experiments

In this section we present numerical experiments illustrating the performance of the row-order Schatten p -norm estimator described in Section 3.2. We simulate the row-order stream by reading the input matrix row by row.

The inputs used are $\{0, 1\}^{n \times n}$ matrices, representing collaboration network graphs (nodes represent scientists and edges represent co-authoring a paper) from the e-print arXiv for scientific collaborations in five different areas in Physics. The data was obtained from the Stanford Large Network Dataset Collection (Leskovec and Krevl, 2014) which was in-turn obtained from (Leskovec et al., 2007). In order to study the effect of sparsity, we “sparsify” each (of five) matrix by sampling 10 nonzero entries in each row uniformly at random (note that max column-sparsity can be larger than 10).

In the first experiment, we use the arXiv General Relativity and Quantum Cosmology collaboration network which has $n = 5242$ rows and columns; after “sparsifying” the matrix as mentioned, the max column-sparsity is 37 and the average sparsity is 6.1. We fix the value of p to be 6, and using our algorithm from Section 3.2, we vary number of estimators (walks) t and compute the *relative error* of the average of the t walks. We repeat this process 10 times for every value of t and plot the mean and standard deviation in Figure 1. In addition, we show in this figure the results of running the same experiment on a “noisy” version of the matrix, by adding to it an error matrix where $1/5$ of the entries are drawn independently from $\mathcal{N}(0, 0.1^2)$ ⁵.

In the second experiment, we use all five collaboration networks – General Relativity and Quantum Cosmology ($n = 5242$), High Energy Physics - Phenomenology ($n = 9877$), High Energy Physics - Theory ($n = 12008$), Astro Physics ($n = 18772$) and Condensed Matter ($n = 23133$). For each matrix we compute walks (estimator from Section 3.2) un-

til the mean of the walks is within 10% of the true Schatten 6-norm of the matrix. We repeat this process 10 times for each matrix and plot the median, the first and third quartile of the number of walks for the 10 trials in Figure 2. Since in the second and third experiments, most of the outputs of the 10 trials are concentrated around the median except for very few trials (one or two) which are very large outliers. Hence, we chose to output the first and third quartiles indicating the output of the majority of the trials.

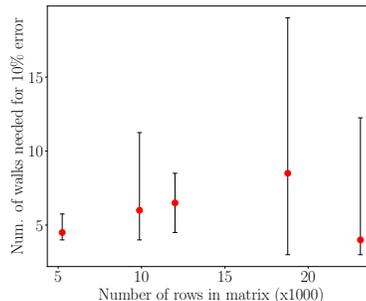


Figure 2: Number of walks to (1 ± 0.1) -approximate Schatten 6-norm of 5 different matrices from arXiv Physics Collaboration Network.

In our third experiment we compute the number of walks needed for the mean of the walks to be within 10% of the true Schatten p -Norm of the GR-QC matrix for different values of p . We vary the value of p and, for each value of p , compute the number of walks needed for 10 trials and plot the median, first and third quartile of the 10 trials in Figure 3.

Our results show that the number of walks, which is a proxy for the amount of space and *time* required to compute our estimator, is much smaller (on the order of 10^2) than the dimension of the matrix (on the order of 10^3). We see that this is indeed true for larger matrices and values of p too.

⁵This value assures the l_2 -norm of the error in a row is “comparable” to the l_2 -norm of the data: $(0.1)^2 \times 5242 \times 0.2 \approx 10 = \text{max row-sparsity}$.

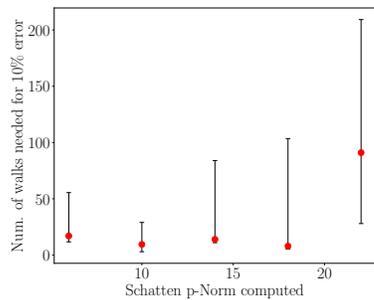


Figure 3: Number of walks to (1 ± 0.1) -approximate Schatten p -norm for arXiv GR-QC Matrix for different values of $p \in 2\mathbb{Z}^+$.

Acknowledgements

The first author was supported by a NSF CAREER grant 1652257, ONR Award N00014-18-1-2364 and and the Lifelong Learning Machines program from DARPA/MTO. The second and last author were partially supported by ONR Award N00014-18-1-2364, the Israel Science Foundation grant #1086/18, and a Minerva Foundation grant. Part of this work was done while some authors were visiting the Simons Institute for the Theory of Computing. The third author was supported by a NSF CAREER grant 1652257 and ONR Award N00014-18-1-2364.

References

- Alexandr Andoni and Huy L. Nguyen. Eigenvalues of a matrix in the streaming model. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013*, pages 1729–1737, 2013. doi:[10.1137/1.9781611973105.124](https://doi.org/10.1137/1.9781611973105.124).
- Christos Boutsidis, Petros Drineas, Prabhanjan Kambadur, Eugenia-Maria Kontopoulou, and Anastasios Zouzias. A randomized algorithm for approximating the log determinant of a symmetric positive definite matrix. *Linear Algebra and its Applications*, 533:95–117, 2017.
- Vladimir Braverman. Approximations of Schatten norms via Taylor expansions. In *International Computer Science Symposium in Russia*, pages 70–79. Springer, 2019.
- Vladimir Braverman, Rafail Ostrovsky, and Gregory Varsinger. Weighted sampling without replacement from data streams. *Information Processing Letters*, 115(12): 923 – 926, 2015. doi:[10.1016/j.ipl.2015.07.007](https://doi.org/10.1016/j.ipl.2015.07.007).
- Vladimir Braverman, Stephen R. Chestnut, Robert Krauthgamer, Yi Li, David P. Woodruff, and Lin Yang. Matrix norms in data streams: Faster, multi-pass and row-order. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 648–657, 2018. URL <http://proceedings.mlr.press/v80/braverman18a.html>.
- Vladimir Braverman, Robert Krauthgamer, Aditya Krishnan, and Roi Sinoff. Schatten norms in matrix streams: Hello sparsity, goodbye dimension. *arXiv preprint arXiv:1907.05457*, 2019.
- Marc Bury and Chris Schwiegelshohn. Sublinear estimation of weighted matchings in dynamic data streams. In *23rd Annual European Symposium, ESA'15*, pages 263–274, 2015. doi:[10.1007/978-3-662-48350-3_23](https://doi.org/10.1007/978-3-662-48350-3_23).
- Jie Chen. How accurately should I compute implicit matrix-vector products when applying the Hutchinson trace estimator? *SIAM J. Scientific Computing*, 38(6), 2016. doi:[10.1137/15M1051506](https://doi.org/10.1137/15M1051506).
- Kenneth L. Clarkson and David P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing, STOC '09*, pages 205–214. ACM, 2009. doi:[10.1145/1536414.1536445](https://doi.org/10.1145/1536414.1536445).
- David Cohen-Steiner, Weihao Kong, Christian Sohler, and Gregory Valiant. Approximating the spectrum of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18*, pages 1263–1271. ACM, 2018. doi:[10.1145/3219819.3220119](https://doi.org/10.1145/3219819.3220119).
- Juri Ganitkevitch, Benjamin van Durme, and Chris Callison-Burch. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764. Association for Computational Linguistics, 2013. URL <https://www.aclweb.org/anthology/N13-1092>.
- Phillip B. Gibbons and Yossi Matias. Synopsis data structures for massive data sets. In *External Memory Algorithms*, volume 50 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, 1999. URL <http://theory.stanford.edu/~matias/papers/synopsis.pdf>.
- N. Goyette, P. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. Changedetection.net: A new change detection benchmark dataset. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2012. doi:[10.1109/CVPRW.2012.6238919](https://doi.org/10.1109/CVPRW.2012.6238919).
- André Gronemeier. Asymptotically optimal lower bounds on the NIH-multi-party information complexity of the

- AND-function and Disjointness. In *26th International Symposium on Theoretical Aspects of Computer Science STACS 2009*, volume 3, pages 505–516. IBFI Schloss Dagstuhl, 2009. doi:[10.4230/lipiCS.stacs.2009.1846](https://doi.org/10.4230/lipiCS.stacs.2009.1846).
- Ivan Gutman, Hanyuan Deng, and Slavko Radenkovic. The Estrada index: An updated survey. *Selected Topics on Applications of Graph Spectra, Math. Inst., Beograd*, pages 155–174, 2011.
- Insu Han, Dmitry Malioutov, and Jinwoo Shin. Large-scale log-determinant computation through stochastic Chebyshev expansions. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 908–917, 2015.
- Insu Han, Dmitry Malioutov, Haim Avron, and Jinwoo Shin. Approximating spectral sums of large-scale matrices using stochastic Chebyshev approximations. *SIAM J. Scientific Computing*, 39(4), 2017. doi:[10.1137/16M1078148](https://doi.org/10.1137/16M1078148).
- T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of AND. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 12th International Workshop, APPROX 2009, and 13th International Workshop, RANDOM 2009*, pages 562–573, 2009. doi:[10.1007/978-3-642-03685-9_42](https://doi.org/10.1007/978-3-642-03685-9_42).
- Ashish Khetan and Sewoong Oh. Spectrum estimation from a few entries. *Journal of Machine Learning Research*, 20(21):1–55, 2019. URL <http://jmlr.org/papers/v20/18-027.html>.
- Weihao Kong and Gregory Valiant. Spectrum estimation from samples. *The Annals of Statistics*, 45, 01 2016. doi:[10.1214/16-AOS1525](https://doi.org/10.1214/16-AOS1525).
- Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. <http://snap.stanford.edu/data>, June 2014.
- Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, 1(1):2?es, March 2007. doi:[10.1145/1217299.1217301](https://doi.org/10.1145/1217299.1217301).
- Yi Li and David P. Woodruff. On approximating functions of the singular values in a stream. In *Proceedings of the 48th annual ACM symposium on Theory of Computing (STOC)*, pages 726–739. ACM, 2016a. doi:[10.1145/2184319.2184343](https://doi.org/10.1145/2184319.2184343).
- Yi Li and David P. Woodruff. Tight bounds for sketching the operator norm, Schatten norms, and subspace embeddings. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016*, pages 39:1–39:11, 2016b. doi:[10.4230/LIPIcs.APPROX-RANDOM.2016.39](https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2016.39).
- Yi Li and David P. Woodruff. Embeddings of Schatten norms with applications to data streams. In *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017*, pages 60:1–60:14, 2017. doi:[10.4230/LIPIcs.ICALP.2017.60](https://doi.org/10.4230/LIPIcs.ICALP.2017.60).
- Yi Li, Huy L. Nguyen, and David P. Woodruff. On sketching matrix norms and the top singular vector. In *Proceedings of the 25th annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1562–1581. SIAM, 2014. doi:[10.1137/1.9781611973402.114](https://doi.org/10.1137/1.9781611973402.114).
- Edo Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '13*, pages 581–588. ACM, 2013. doi:[10.1145/2487575.2487623](https://doi.org/10.1145/2487575.2487623).
- X. Liu, G. Zhao, J. Yao, and C. Qi. Background subtraction based on low-rank and structured sparse decomposition. *IEEE Transactions on Image Processing*, 24(8): 2502–2514, 2015. doi:[10.1109/TIP.2015.2419084](https://doi.org/10.1109/TIP.2015.2419084).
- Morteza Monemizadeh and David P. Woodruff. 1 pass relative-error L_p -sampling with applications. In *Proceedings of the Twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms, SODA '10*, pages 1143–1160. SIAM, 2010. URL <http://dl.acm.org/citation.cfm?id=1873601.1873693>.
- Cameron Musco, Praneeth Netrapalli, Aaron Sidford, Shashanka Ubaru, and David P. Woodruff. Spectrum approximation beyond fast matrix multiplication: Algorithms and hardness. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 8:1–8:21. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018. doi:[10.4230/LIPIcs.ITCS.2018.8](https://doi.org/10.4230/LIPIcs.ITCS.2018.8).
- Edoardo Di Napoli, Eric Polizzi, and Yousef Saad. Efficient estimation of eigenvalue counts in an interval. *Numerical Lin. Alg. with Applications.*, 23(4):674–692, 2016. doi:[10.1002/nla.2048](https://doi.org/10.1002/nla.2048).
- Shashanka Ubaru and Yousef Saad. Applications of trace estimation techniques. In *High Performance Computing in Science and Engineering - Third International Conference, HPCSE 2017, Karolinka, Czech Republic, May 22-25, 2017, Revised Selected Papers*, pages 19–33, 2017. doi:[10.1007/978-3-319-97136-0_2](https://doi.org/10.1007/978-3-319-97136-0_2).

Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of $\text{tr}(f(a))$ via stochastic Lanczos quadrature. *SIAM J. Matrix Analysis Applications*, 38(4):1075–1099, 2017. doi:[10.1137/16M1104974](https://doi.org/10.1137/16M1104974).

Elad Verbin and Wei Yu. The streaming complexity of cycle counting, sorting by reversals, and other problems. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 11–25, 2011. URL <http://dl.acm.org/citation.cfm?id=2133036.2133038>.

Jeffrey Scott Vitter. Random sampling with a reservoir. *ACM Trans. Math. Softw.*, 11(1):37–57, 1985. doi:[10.1145/3147.3165](https://doi.org/10.1145/3147.3165).

Jeffrey Scott Vitter. External memory algorithms and data structures: Dealing with massive data. *ACM Computing Survey.*, 33(2):209–271, 2001. doi:[10.1145/384192.384193](https://doi.org/10.1145/384192.384193).

Lingfei Wu, Jesse Laeuchli, Vassilis Kalantzis, Andreas Stathopoulos, and Efstratios Gallopoulos. Estimating the trace of the matrix inverse by interpolating from the diagonal of an approximate inverse. *J. Comput. Physics*, 326:828–844, 2016. doi:[10.1016/j.jcp.2016.09.001](https://doi.org/10.1016/j.jcp.2016.09.001).

Yuan Xie, Shuhang Gu, Yan Liu, Wangmeng Zuo, Wensheng Zhang, and Lei Zhang. Weighted Schatten p -norm minimization for image denoising and background subtraction. *IEEE Transactions on Image Processing*, 25:4842–4857, 2016.

Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan. Distributed estimation of generalized matrix rank: Efficient algorithms and lower bounds. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15*, pages 457–465. JMLR.org, 2015. URL <http://dl.acm.org/citation.cfm?id=3045118.3045168>.