Single Pass Spectral Sparsification in Dynamic Streams

Michael Kapralov
MIT
kapralov@mit.edu

Yin Tat Lee
MIT
yintat@mit.edu

Cameron Musco
MIT
cnmusco@mit.edu

Christopher Musco
MIT
cpmusco@mit.edu

Aaron Sidford
MIT
sidford@mit.edu

Abstract

We present the first single pass algorithm for computing spectral sparsifiers of graphs in the dynamic semi-streaming model. Given a single pass over a stream containing insertions and deletions of edges to a graph G, our algorithm maintains a randomized linear sketch of the incidence matrix of G into dimension $O(\frac{1}{\epsilon^2} n \operatorname{polylog}(n))$. Using this sketch, at any point, the algorithm can output a $(1 \pm \epsilon)$ spectral sparsifier for G with high probability.

While $O(\frac{1}{\epsilon^2} n \operatorname{polylog}(n))$ space algorithms are known for computing *cut sparsifiers* in dynamic streams [AGM12b, GKP12] and spectral sparsifiers in *insertion-only* streams [KL11], prior to our work, the best known single pass algorithm for maintaining spectral sparsifiers in dynamic streams required sketches of dimension $\Omega(\frac{1}{\epsilon^2} n^{5/3})$ [AGM14].

To achieve our result, we show that, using a coarse sparsifier of G and a linear sketch of G's incidence matrix, it is possible to sample edges by effective resistance, obtaining a spectral sparsifier of arbitrary precision. Sampling from the sketch requires a novel application of ℓ_2/ℓ_2 sparse recovery, a natural extension of the ℓ_0 methods used for cut sparsifiers in [AGM12b]. Recent work of [MP12] on row sampling for matrix approximation gives a recursive approach for obtaining the required coarse sparsifiers.

Under certain restrictions, our approach also extends to the problem of maintaining a spectral approximation for a general matrix $A^\top A$ given a stream of updates to rows in A.

1 Introduction

1.1 The Dynamic Semi-Streaming Model

When processing massive graph datasets arising from social networks, web topologies, or interaction graphs, computation may be as limited by space as it is by runtime. To cope with this issue, one might hope to apply techniques from the streaming model of computation, which restricts algorithms to few passes over the input and space polylogarithmic in the input size. Streaming algorithms have been studied extensively in various application domains – see [Mut05] for an overview. However, the model has proven too restrictive for even the simplest graph algorithms. For example, testing $s-t$ connectivity requires $\Omega(n)$ space [HRR99].

The less restrictive semi-streaming model, in which the algorithm is allowed $\tilde{O}(n)$ space, is more suited for graph algorithms [FKM+05], and has received significant attention in recent years. In this model, a processor receives a stream of edges over a fixed set of n nodes. Ideally, the processor should only have to perform a single pass (or few passes) over the edge stream, and the processing time per edge, as well as the time required to output the final answer, should be small.

In the *dynamic semi-streaming model*, the graph stream may include both edge insertions and deletions [AGM12a]. This extension captures the fact that large graphs are unlikely to be static. Dynamic semi-streaming algorithms allow us to quickly process general updates in the form of edge insertions and deletions to maintain a small-space representation of the graph from which we can later compute a result. Sometimes the dynamic model is referred to as the *insertion-deletion model*, in contrast to the more restrictive *insertion-only model*.

Work on semi-streaming algorithms in both the dynamic and insertion-only settings is extensive. Researchers have tackled connectivity, bipartiteness, minimum spanning trees, maximal matchings, and spanners among other problems [FKM+05, ELMS11, Elk11, AGM12a, AGM12b]. In [McG13], McGregor surveys much of this progress and provides a more complete list of citations.

1.2 Streaming Sparsification

First introduced by Benczúr and Karger [BK96], a *cut sparsifier* of a graph G is a weighted subgraph with only $O(\frac{1}{\epsilon^2} n \operatorname{polylog}(n))$ edges that preserves the total edge weight over every cut in G to within a $(1 \pm \epsilon)$ multiplicative factor. Cut sparsifiers can be used to compute approximations for minimum cut, sparsest cut, maximum flow, and a variety of other problems over G. In [ST04], Spielman and Teng introduce the stronger *spectral sparsifier*, a weighted subgraph whose Laplacian spectrally approximates the Laplacian of G. In addition to maintaining the cut approximation of Benczúr and Karger, spectral sparsifiers can be used to approximately solve linear systems over the Laplacian of G, and to approximate effective resistances, spectral clusterings, random walk properties, and a variety of other computations.

The problem of computing graph sparsifiers in the streaming model has received a lot of attention. Ahn and Guha give the first single pass, insertion-only algorithm for cut sparsifiers [AG09]. Kelner and Levin give a single pass, insertion-only algorithm for spectral sparsifiers [KL11]. This algorithm stores a sparse graph: edges are added as they are streamed in and, when the graph grows too large, it is resparsified. The construction is very clean, but inherently does not extend to the dynamic model since, to handle edge deletions, we need more information than just a sparsifier itself. Edges eliminated to create an intermediate sparsifier may become critically important later if other edges are deleted, so we need to maintain information that allows recovery of such edges.

Ahn, Guha, and McGregor make a very important insight in [AGM12a], demonstrating the power of linear graph sketches in the dynamic model. They present the first dynamic algorithm for cut sparsifiers, which initially required $O(\frac{1}{\epsilon^2}n^{1+\gamma})$ space and $O(1/\gamma)$ passes over the graph stream. However, the result was later improved to a single pass and $O(\frac{1}{\epsilon^2}n\,\text{polylog}(n))$ space [AGM12b, GKP12]. Our algorithm extends the sketching and sampling approaches from these papers to the spectral problem.

In [AGM13], the authors show that linear graph sketches that capture connectivity information can be used to coarsely approximate spectral properties and they obtain spectral sparsifiers using $O(\frac{1}{\epsilon^2}n^{5/3}\,\text{polylog}(n))$ space in the dynamic setting. However, they also show that their coarse approximations are tight, so a new approach is required to obtain spectral sparsifiers using just $O(\frac{1}{\epsilon^2}n\,\text{polylog}(n))$ space. They conjecture that a dynamic algorithm for doing so exists. The development of such an algorithm is also posed as an open question in [McG13]. A two-pass algorithm for constructing a spectral sparsifier in the dynamic streaming model using $O\left(\frac{1}{\epsilon^2}n^{1+o(1)}\right)$ space is presented in [KW14]. Their approach is very different from ours: it leverages a reduction from spanner constructions to spectral sparsification presented in [KP12]. It is not known if this approach extends to a space efficient single pass algorithm.

1.3 Our Contribution

Our main result is an algorithm for maintaining a small graph sketch from which we can recover a spectral sparsifier. For simplicity, we present the algorithm in the case of unweighted graphs. However, in Section 6, we show that it is easily extended to weighted graphs, as long as an edge's weight is specified when it is deleted. This model matches what is standard for dynamic cut sparsifiers [AGM12b, GKP12].

Theorem 1 (Main Result)**.** *There exists an algorithm that, for any $\epsilon > 0$, processes a list of edge insertions and deletions for an unweighted graph G in a single pass and maintains a set of linear sketches of this input in $O\left(\frac{1}{\epsilon^2}n\,\text{polylog}(n)\right)$ space. From these sketches, it is possible to recover, with high probability, a weighted subgraph H with $O(\frac{1}{\epsilon^2}n\log n)$ edges such that H is a $(1\pm\epsilon)$ spectral sparsifier of G. The algorithm recovers H in $O\left(\frac{1}{\epsilon^2}n^2\,\text{polylog}(n)\right)$ time.*

It is well known that independently sampling edges from a graph G according to their *effective resistances* gives a $(1\pm\epsilon)$ spectral sparsifier of G with $O(\frac{1}{\epsilon^2}n\log n)$ edges [SS08]. We can 'refine' any coarse sparsifier for G by using it to approximate effective resistances and then resample edges according to these approximate resistances. We show how to perform this refinement in the streaming setting, extending graph sketching techniques initially used for cut sparsifiers ([AGM12b, GKP12]) and introducing a new sampling technique based on an ℓ_2 heavy hitters algorithm. Our refinement procedure is combined with a clever recursive method for obtaining a coarse sparsifier introduced by Miller and Peng in a preprint of a recent paper on iterative row sampling for matrix approximation [MP12].

The fact that our algorithm maintains a linear sketch of the streamed graph allows for the simple handling of edge deletions, which are treated as negative edge insertions. Additionally, due to their linearity, our sketches are composable - sketches of subgraphs can simply be added to produce a sketch of the full graph. Thus, our techniques are directly applicable in distributed settings where separate processors hold different subgraphs or each processes different edge substreams.

Our application of linear sketching also gives a nice information theoretic result on graph compression. A spectral sparsifier is a powerful compression for a graph. It maintains, up to an ϵ

factor, all spectral information about the Laplacian using just $O(\frac{1}{\epsilon^2} n \log n)$ space. At first glance, it may seem that such a compression requires careful analysis of the input graph to determine what information to keep and what to discard. However, the non-adaptive linear sketches used in our algorithm are completely *oblivious*: at each edge insertion or deletion, we do not need to examine the current compression at all to make the appropriate update. As in sparse recovery or dimensionality reduction, we essentially just multiply the vertex edge incidence matrix by a random projection matrix, decreasing its height drastically in the process. Nevertheless, the oblivious compression obtained holds as much information as a spectral sparsifier - in fact, we show how to extract a spectral sparsifier from it! Furthermore, the compression is only larger than $O(\frac{1}{\epsilon^2} n \log n)$ by log factors. Our result is the first of this kind in the spectral domain. The only other streaming algorithm for spectral sparsification that uses $O(\frac{1}{\epsilon^2} n \operatorname{polylog}(n))$ space is distinctly non-oblivious [KL11] and oblivious subspace embeddings for compressing general matrices inherently require $O(n^2 \operatorname{polylog}(n))$ space, even when the matrix is sparse (as in the case of an edge vertex incidence matrix) [Sar06, NN13].

Finally, it can be noted that our proofs rely very little on the fact that our data stream represents a graph. We show that, with a few modifications, given a stream of row updates for a general structured matrix A, it is possible to maintain a $O(\frac{1}{\epsilon^2} n \operatorname{polylog}(n))$ sized sketch from which a spectral approximation to $A^\top A$ can be recovered. By structured, we mean any matrix whose rows are selected from some fixed dictionary of size $\operatorname{poly}(n)$. Spectral graph sparsification is a special case of this problem: set A to be the vertex edge incidence matrix of our graph. The dictionary is the set of all possible $\binom{n}{2}$ edge rows that may appear in A and $A^\top A$ is the graph Laplacian.

1.4 Road Map

Section 2 Lay out notation, build linear algebraic foundations for spectral sparsification, and present lemmas for graph sampling and sparse recovery required by our algorithm.

Section 3 Give an overview of our central algorithm, providing intuition and motivation.

Section 4 Present an algorithm of Miller and Peng ([MP12]) for building a chain of coarse sparsifiers and prove our main result, assuming a primitive for sampling edges by effective resistance in the streaming model.

Section 5 Develop this sampling primitive, our main technical contribution.

Section 6 Show how to extend the algorithm to weighted graphs.

Section 7 Show how to extend the algorithm to general structured matrices.

Section 8 Remove our assumption of fully independent hash functions, using a pseudorandom number generator to achieve a final small space algorithm.

2 Notation and Preliminaries

2.1 Graph Notation

Let $\mathbf{B}_n \in \mathbb{R}^{\binom{n}{2} \times n}$ be the vertex edge incidence matrix of the undirected, unweighted complete graph over n vertices. $\mathbf{b}_e$, the row corresponding to edge $e = (u, v)$ contains a 1 in column u, a (-1) in column v, and 0's elsewhere.

We write the vertex edge incidence matrix of any unweighted, undirected graph $G(V, E)$ as $\mathbf{B} = \mathbf{S}\mathbf{B}_n$ where $\mathbf{S}$ is an $\binom{n}{2} \times \binom{n}{2}$ diagonal matrix with ones at positions corresponding to edges contained in G and zeros elsewhere.[1] The $n \times n$ Laplacian matrix of G is given by $\mathbf{K} = \mathbf{B}^\top \mathbf{B}$.

2.2 Spectral Sparsification

For any matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$, $\tilde{\mathbf{K}}$ is a $(1 \pm \epsilon)$ spectral sparsifier of $\mathbf{K} = \mathbf{B}^\top \mathbf{B}$ if, $\forall \mathbf{x} \in \mathbb{R}^n$, $(1-\epsilon)\mathbf{x}^\top \mathbf{K}\mathbf{x} \leq \mathbf{x}^\top \tilde{\mathbf{K}}\mathbf{x} \leq (1+\epsilon)\mathbf{x}^\top \mathbf{K}\mathbf{x}$. This condition can also be written as $(1-\epsilon)\mathbf{K} \preceq \tilde{\mathbf{K}} \preceq (1+\epsilon)\mathbf{K}$ where $\mathbf{C} \preceq \mathbf{D}$ indicates that $\mathbf{D} - \mathbf{C}$ is positive semidefinite. More succinctly, $\tilde{\mathbf{K}} \approx_\epsilon \mathbf{K}$ denotes the same condition. We'll also use the slightly weaker notation $(1-\epsilon)\mathbf{K} \preceq_r \tilde{\mathbf{K}} \preceq_r (1+\epsilon)\mathbf{K}$ to indicate that $(1-\epsilon)\mathbf{x}^\top \mathbf{K}\mathbf{x} \leq \mathbf{x}^\top \tilde{\mathbf{K}}\mathbf{x} \leq (1+\epsilon)\mathbf{x}^\top \mathbf{K}\mathbf{x}$ for all x in the *row span* of $\mathbf{K}$ (which is the same as the row span of $\mathbf{B}$). If $\tilde{\mathbf{K}}$ has the same row span as $\mathbf{K}$ this notation is equivalent to the initial notion of spectral sparsification.

Note that we are giving these definitions for a general matrix $\mathbf{B}$, however well will often work with the case where $\mathbf{B}$ is the vertex edge incidence matrix of a graph G and $\mathbf{K}$ is the graph Laplacian. We will not always require our approximation $\tilde{\mathbf{K}}$ to be the graph Laplacian of a weighted subgraph, which is a standard assumption. For this reason, we avoid the standard $\mathbf{L}_G$ notation for the Laplacian. For our purposes, $\tilde{\mathbf{K}}$ will always be a sparse symmetric diagonally dominant matrix, containing no more than $O(n \log n)$ non-zero entries. In fact, it will always be the Laplacian of a sparse subgraph, but possibly with weight added to its diagonal entries. Furthermore, the final approximation returned by our streaming algorithm will be a bonafide spectral graph sparsifier - the Laplacian matrix of a weighted subgraph of G.

2.3 Leverage Scores and Row Sampling

For any $\mathbf{B} \in \mathbb{R}^{m \times n}$ with rank r, let $\mathbf{K}^+$ denote the Moore-Penrose pseudoinverse of $\mathbf{K} = \mathbf{B}^\top \mathbf{B}$. Consider the reduced singular value decomposition, $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. $\mathbf{U} \in \mathbb{R}^{m \times r}$ and $\mathbf{V} \in \mathbb{R}^{n \times r}$ have orthonormal columns and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is diagonal and contains the nonzero singular values of $\mathbf{B}$. $\mathbf{K} = \mathbf{B}^\top \mathbf{B} = \mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^\top$. It follows that:

$$\mathbf{K}^+ = \mathbf{V}(\mathbf{\Sigma}^{-1})^2\mathbf{V}^\top$$

The leverage score, τ_i, for a row $\mathbf{b_i}$ in $\mathbf{B}$ is defined as:

$$\tau_i \stackrel{\text{def}}{=} \mathbf{b}_i^\top \mathbf{K}^+ \mathbf{b}_i = \mathbf{u}_i^\top \mathbf{\Sigma}\mathbf{V}^\top (\mathbf{V}\mathbf{\Sigma}^{-2}\mathbf{V}^\top)\mathbf{V}\mathbf{\Sigma}\mathbf{u}_i = \|\mathbf{u}_i\|_2^2 \leq 1$$

The last inequality follows from the fact that every row in a matrix with orthonormal columns has norm less than 1. In a graph, $\tau_i = r_i w_i$, where r_i is the *effective resistance* of edge i and w_i is its weight. Furthermore:

$$\sum_{i=1}^{m} \tau_i = \text{tr}(\mathbf{B}\mathbf{K}^+\mathbf{B}^\top) = \|\mathbf{U}\|_F^2 = r = \text{rank}(\mathbf{B})$$

It is well known that by sampling the rows of $\mathbf{B}$ according to their leverage scores it is possible to obtain a matrix $\tilde{\mathbf{B}}$ such that $\tilde{\mathbf{K}} = \tilde{\mathbf{B}}^\top \tilde{\mathbf{B}} \approx_\epsilon \mathbf{K}$ with high probability. Furthermore, if obtaining exact

[1] Typically the rows of $\mathbf{B}$ that are all 0 are removed, however we find this formulation more convenient for our purposes.

leverage scores is computationally difficult, it suffices to sample by upper bounds on the scores. Typically, rows are sampled with replacement with probability proportional to their leverage score [SS08, LMP13]. We give an alternative independent sampling procedure based off the matrix concentration results of [Tro12], which is more amenable to our application.

Lemma 1 (Spectral Sparsifier via Leverage Score Sampling)**.** *Let $\tilde{\boldsymbol{\tau}}$ be a vector of m estimated leverage scores for the rows of $\mathbf{B}$, such that $1 \geq \tilde{\tau}_i \geq \tau_i$ for all $i \in [m]$. For some known constant c, let $\mathbf{W}_1, \mathbf{W}_2, ...\mathbf{W}_{c\log n\epsilon^{-2}}$ be diagonal matrices with independently chosen entries such that $\mathbf{W}_j(i,i) = \frac{1}{\tilde{\tau}_i}$ with probability $\tilde{\tau}_i$ and $\mathbf{W}_j(i,i) = 0$ otherwise. Letting $\bar{\mathbf{W}} = \frac{1}{c\log n\epsilon^{-2}} \cdot \sum_j \mathbf{W}_j$ then with high probability,*

$$\tilde{\mathbf{K}} = \mathbf{B}^\top \bar{\mathbf{W}} \mathbf{B} \approx_\epsilon \mathbf{K}$$

Furthermore, $\bar{\mathbf{W}}$ has $O(\|\tilde{\boldsymbol{\tau}}\|_1 \log n\epsilon^{-2})$ nonzeros with high probability. That is, if we sample each each row of $\mathbf{B}$ independently with probability $\tilde{\tau}_i$, reweight selected rows by $\frac{1}{\sqrt{\tilde{\tau}_i}}$, and average over $c\log n\epsilon^{-2}$ trials, we obtain a matrix $\tilde{\mathbf{B}} = \mathbf{B}\bar{\mathbf{W}}^{1/2}$ such that $\tilde{\mathbf{B}}^\top\tilde{\mathbf{B}} = \mathbf{B}^\top\bar{\mathbf{W}}\mathbf{B} = \tilde{\mathbf{K}} \approx_\epsilon \mathbf{K}$ and $\tilde{\mathbf{B}}$ contains just $O(\|\tilde{\boldsymbol{\tau}}\|_1 \log n\epsilon^{-2})$ reweighted rows of $\mathbf{B}$ with high probability.

We relegate the proof of Lemma 1 to Appendix A.

2.4 Sparse Recovery

We now give a sparse recovery primitive that is used to sample edges from our linear sketches. We use an ℓ_2 heavy hitters algorithm that, for any vector $\mathbf{x}$, lets us recover from a small size sketch $\mathbf{\Phi x}$, the index i and the approximate value of $\mathbf{x}_i$ for all i such that $\mathbf{x}_i > \frac{1}{O(\text{polylog}(n))}||\mathbf{x}||_2$.

Lemma 2 (ℓ_2 Heavy Hitters)**.** *For each $\eta > 0$, there is a decoding algorithm D and a distribution on matrices $\mathbf{\Phi}$ in $\mathbb{R}^{O(\eta^{-2}\,\text{polylog}(N))\times N}$ such that, for any $\mathbf{x} \in \mathbb{R}^N$, with probability $1 - N^{-c}$ over the choice of $\mathbf{\Phi}$, given $\mathbf{\Phi x}$, the algorithm D returns a vector $\mathbf{w}$ such that $\mathbf{w}$ has $O(\eta^{-2}\,\text{polylog}(N))$ non-zeros and satisfies*

$$||\mathbf{x} - \mathbf{w}||_\infty \leq \eta||\mathbf{x}||_2.$$

with probability $1 - N^{-c}$. The sketch $\mathbf{\Phi x}$ can be maintained and decoded in $O(\eta^{-2}\,\text{polylog}(N))$ space.

Note that setting $\eta = \frac{\epsilon}{C\log n}$ for any $0 < \epsilon < 1/2$ and $C > 0$, guarantees that $\mathbf{w}_i$ must be a $(1\pm\epsilon)$ approximation of $\mathbf{x}_i$ for any i with $\mathbf{x}_i \geq \frac{1}{C\log n}\|\mathbf{x}\|_2$. It also guarantees that we can distinguish using $\mathbf{w}_i$ whether $\mathbf{x}_i \geq \frac{1}{C\log n}\|\mathbf{x}\|_2$ or $\mathbf{x}_i < \frac{1}{2C\log n}\|\mathbf{x}\|_2$. We give a proof of Lemma 2 in Appendix B

3 Algorithm Overview

Before providing a formal presentation and proof of our main result, Theorem 1, we would like to give an informal overview of the algorithm to provide intuition.

3.1 Effective Resistances

As explained in Section 2.3, spectral sparsifiers can be generated by sampling edges, i.e. rows of the vertex edge incidence matrix. For an unweighted graph G, each edge is sampled independently with probability equal to its leverage score, τ_e. After appropriate repetition of the sampling, we reweight and combine any sampled edges. The result is a subgraph of G containing, with high probability, $O(\frac{1}{\epsilon^2} n \log n)$ edges and spectrally approximating G.

If we view G as an electrical circuit, with each edge representing a unit resistor, the leverage score of an edge $e = (i, j)$ is equivalent to its effective resistance. This value can be computed by forcing 1 unit of current out of vertex i and 1 unit of current into vertex j. The resulting voltage difference between the two vertices is the effective resistance of e. Qualitatively, if the voltage drop is low, there are many low resistance (i.e. short) paths between i and j. Thus, maintaining a direct connection between these vertices is less critical in approximating G, so e is less likely to be sampled. Effective resistance can be computed as:

$$\tau_e = \mathbf{b}_e^\top \mathbf{K}^+ \mathbf{b}_e$$

Note that τ_e can be computed for any pair of vertices, (i, j), or in other words, for any possible edge in G. We can evaluate $\mathbf{b}_e^\top \mathbf{K}^+ \mathbf{b}_e$ even if e is not present in the graph. Thus, we can reframe our sampling procedure. Instead of just sampling edges actually in G, imagine we run a sampling procedure for *every possible* e. When recombining edges to form a spectral sparsifier, we separately check whether each edge e is in G and only insert into the sparsifier if it is.

3.2 Sampling in the Streaming Model

With this procedure in mind, a sampling method that works in the streaming setting requires two components. First, we need to obtain a constant factor approximation to τ_e for any e. Known sampling algorithms, including our Lemma 1, are robust to this level of estimation. Second, we need to compress our edge insertions and deletions in such a way that, during post-processing of our sketch, we can determine whether or not a sampled edge e actually exists in G.

The first requirement is achieved through the recursive procedure given in [MP12]. We will give the overview shortly but, for now, assume that we have access to a coarse sparsifier, $\tilde{\mathbf{K}} \approx_{1/2} \mathbf{K}$. Computing $\mathbf{b}_e^\top \tilde{\mathbf{K}}^+ \mathbf{b}_e$ gives a 2 factor multiplicative approximation of τ_e for each e. Furthermore, as long as $\tilde{\mathbf{K}}$ has sparsity $O(n \operatorname{polylog}(n))$, the computation can be done in small space using any nearly linear time solver for symmetric diagonally dominant linear systems (e.g. [KMP11]).

Solving part two (determining which edges are actually in G) is a bit more involved. As a first step, consider writing:

$$\tau_e = \mathbf{b}_e^\top \mathbf{K}^+ \mathbf{K} \mathbf{K}^+ \mathbf{b}_e = \|\mathbf{B}\mathbf{K}^+ \mathbf{b}_e\|_2^2 = \|\mathbf{S}\mathbf{B}_n \mathbf{K}^+ \mathbf{b}_e\|_2^2$$

Referring to Section 2, recall that $\mathbf{B} = \mathbf{S}\mathbf{B}_n$ is exactly the same as a standard vertex edge incidence matrix except that rows in $\mathbf{B}_n$ corresponding to nonexistent edges are zeroed out instead of removed. Denote $\mathbf{x}_e = \mathbf{S}\mathbf{B}_n \mathbf{K}^+ \mathbf{b}_e$. Each nonzero entry in $\mathbf{x}_e$ contains the voltage difference across some edge (resistor) in G when one unit of current is forced from i to j.

When e is not in G, the e^{th} entry of $\mathbf{x}_e$, $\mathbf{x}_e(e)$, is 0. However, if e is in G, then $\mathbf{x}_e(e)$ is τ_e. Furthermore, $\|\mathbf{x}_e\|_2^2 = \tau_e$. So, if we could access a sketch of $\mathbf{x}_e$, could we determine whether or not $e \in G$ using our ℓ_2 sparse recovery primitive?

Not quite - to determine whether an index in $\mathbf{x}_e$ is nonzero, the recovery primitive, Lemma 2, requires it to account for an $O(1/\operatorname{polylog}(n))$ fraction of the total ℓ_2 norm. Currently, $\mathbf{x}_e(e)/\|\mathbf{x}_e\|_2 = \sqrt{\tau_e}$, which could be much smaller than $O(1/\log n)$. However, suppose we had a sketch of $\mathbf{x}_e$ with all but τ_e fraction of edges randomly sampled out. Then, we would expect $\|\mathbf{x}_e\|_2^2 \approx \tau_e^2$ and, in fact, we can show that it would equal $o(\tau_e \log n)$ with high probability. Thus, $\mathbf{x}_e(e)/\|\mathbf{x}_e\|_2 = \Omega(1/\operatorname{polylog}(n))$ and sparse recovery would successfully indicate whether or not $e \in G$. What's more, randomly sampling zeroing out edges of $\mathbf{x}_e$ can serve as our main sampling routine for edge e. This process will set $\mathbf{x}_e(e) = 0$ with probability $(1 - \tau_e)$, exactly what we wanted to sample by in the first place!

However, how do we go about sketching every appropriately sampled $\mathbf{x}_e$? Well, consider subsampling our graph at geometrically decreasing rates, $1/2^s$ for $s \in \{0, 1, ...O(\log n)\}$. Maintain linear sketches $\mathbf{\Pi}_1\mathbf{B}_1, ...\mathbf{\Pi}_{O(\log n)}\mathbf{B}_{O(\log n)}$ of the vertex edge incidence matrix for every subsampled graph using the ℓ_2 sparse recovery sketch distribution from Lemma 2. When asked to output a spectral sparsifier, for every possible edge e, we compute its approximate effective resistance τ_e using $\tilde{\mathbf{K}}$ and determine a rate $1/2^s$ that approximates τ_e.

Next, since our sketches are linear, for every edge, we can just multiply $\mathbf{\Pi}_{1/2^s}\mathbf{B}_{1/2^s}$ on the right by $\tilde{\mathbf{K}}^+\mathbf{b}_e$. We get:

$$\mathbf{\Pi}_{1/2^s}\mathbf{B}_{1/2^s}\tilde{\mathbf{K}}^+\mathbf{b}_e \approx \mathbf{\Pi}_{1/2^s}\mathbf{x}_e^{1/2^s}$$

where $\mathbf{x}_e^{1/2^s}(e)$ is $\mathbf{x}_e$ sampled at rate $1/2^s \approx \tau_e$. This sketch is equivalent to what would be obtained if we had been able to sketch $\mathbf{x}_e^{1/2^s}$ in the first place. Thus, as explained, we can just use our sparse recovery routine to determine whether or not e is present. If it is, we have obtained a sample for our spectral sparsifier!

3.3 A Chain of Coarse Sparsifiers

The final required component is access to some sparse $\tilde{\mathbf{K}} \approx_{1/2} \mathbf{K}$. This coarse sparsifier is obtained recursively by constructing a chain of matrices, $\left[\mathbf{K}(0), \mathbf{K}(1), \ldots, \mathbf{K}(d), \mathbf{K}\right]$ each weakly approximating the next. Specifically, imagine producing $\mathbf{K}(d)$ by adding a fairly light identity matrix to $\mathbf{K}$. As long as the identity's weight is small compared to $\mathbf{K}$'s spectrum, $\mathbf{K}(d)$ approximates $\mathbf{K}$. Add even more weight to the diagonal to form $\mathbf{K}(d-1)$. Again, as long as the increase is small, $\mathbf{K}(d-1)$ approximates $\mathbf{K}(d)$. We continue down the chain until $\mathbf{K}(0)$, which will actually have a heavy diagonal after all the incremental increases. Thus, $\mathbf{K}(0)$ can be approximated by an appropriately scaled identity matrix, which is clearly sparse. Miller and Peng show that parameters can be chosen such that $d = O(\log n)$ [MP12].

Putting everything together, we maintain $O(\log n)$ sketches for $\left[\mathbf{K}(0), \mathbf{K}(1), \ldots, \mathbf{K}(d), \mathbf{K}\right]$. We first use a weighted identity matrix as a coarse approximation for $\mathbf{K}(0)$, which allows us to recover a good approximation to $\mathbf{K}(0)$ from our sketch. This approximation will in turn be a coarse approximation for $\mathbf{K}(1)$, so we can recover a good sparsifier of $\mathbf{K}(1)$. Continuing up the chain, we eventually recover a good sparsifier for our final matrix, $\mathbf{K}$. This approach is formalized in the next section.

4 Recursive Sparsifier Construction

In this section, we describe the recursive procedure for obtaining a chain of coarse sparsifiers using a technique introduced by Miller and Peng - "Introduction and Removal of Artificial Bases" [MP12]. We then formally prove Theorem 1 by combining this technique with the sampling algorithm developed in Section 5.

Theorem 2 (Recursive Sparsification ([MP12], Section 4))**.** *Consider any PSD matrix* $\mathbf{K}$ *with maximum eigenvalue bounded from above by* λ_u *and minimum nonzero eigenvalue bounded from below by* λ_l*. Let* $d = \lceil \log_2(\lambda_u/\lambda_l) \rceil$*. For* $\ell \in \{0, 1, 2, ..., d\}$*, define:*

$$\gamma(\ell) = \lambda_u / 2^\ell$$

So, $\gamma(d) \leq \lambda_l$ *and* $\gamma(0) = \lambda_u$*. Then the chain of PSD matrices,* $\left[\mathbf{K}(0), \mathbf{K}(1), \ldots, \mathbf{K}(d)\right]$ *with:*

$$\mathbf{K}(\ell) = \mathbf{K} + \gamma(\ell)\mathbf{I}_{n \times n}$$

satisfies the following relations:

1. $\mathbf{K} \preceq_r \mathbf{K}(d) \preceq_r 2\mathbf{K}$
2. $\mathbf{K}(\ell) \preceq \mathbf{K}(\ell - 1) \preceq 2\mathbf{K}(\ell)$ *for all* $\ell \in \{1, \ldots, d\}$
3. $\mathbf{K}(0) \preceq 2\gamma(0)\mathbf{I} \preceq 2\mathbf{K}(0)$

When $\mathbf{K}$ *is the Laplacian of an unweighted graph,* $\lambda_{max} < 2n$ *and* $\lambda_{min} > 8/n^2$ *(where here* λ_{min} *is the smallest* nonzero *eigenvalue). Thus the length of our chain,* $d = \lceil \log_2 \lambda_u/\lambda_l \rceil$*, is* $O(\log n)$*.*

For completeness, we've included a proof of Theorem 2 in Appendix [**?**]. Now, to prove our main result, we need to state the sampling primitive for streams that will be developed in Section 5. This procedure maintains a linear sketch of a vertex edge incidence matrix $\mathbf{B}$, and using a coarse sparsifier of $\mathbf{K}(\ell) = \mathbf{B}^\top\mathbf{B} + \gamma(\ell)\mathbf{I}$, performs independent edge sampling as required by Lemma 1, to obtain a better sparsifier of $\mathbf{K}(\ell)$.

Theorem 3. *Let* $\mathbf{B} \in \mathbb{R}^{n \times m}$ *be the vertex edge incidence matrix of an unweighted graph* G*, specified by an insertion-deletion graph stream. Let* $\gamma = O(\text{poly}\, n)$ *be a fixed parameter and consider* $\mathbf{K} = \mathbf{B}^\top\mathbf{B} + \gamma\mathbf{I}$*. For any* $0 < \epsilon < 1$*, there exists a sketching procedure* `MaintainSketches`$(\mathbf{B}, \epsilon)$ *that outputs an* $O(n \,\text{polylog}(n))$ *sized sketch* $\mathbf{\Pi B}$*. There exists a corresponding recovery algorithm* `RefineSparsifier`*, such that, if* $\tilde{\mathbf{K}}$ *is a spectral approximation to* $\mathbf{K}$ *with* $O(n \,\text{polylog}(n))$ *nonzeros and* $c\mathbf{K} \preceq_r \tilde{\mathbf{K}} \preceq_r \mathbf{K}$ *for some constant* $0 < c < 1$ *then:*

`RefineSparsifier`$(\mathbf{\Pi B}, \tilde{\mathbf{K}}, \gamma, \epsilon, c)$ *returns, with high probability,* $\tilde{\mathbf{K}}_\epsilon = \tilde{\mathbf{B}}_\epsilon^\top\tilde{\mathbf{B}}_\epsilon + \gamma\mathbf{I}$*, where* $(1 - \epsilon)\mathbf{K} \preceq_r \tilde{\mathbf{K}}_\epsilon \preceq_r (1 + \epsilon)\mathbf{K}$*, and* $\tilde{\mathbf{B}}_\epsilon$ *contains only* $O(\epsilon^{-2}c^{-1}n \log n)$ *reweighted rows of* $\mathbf{B}$ *with high probability.* `RefineSparsifier` *runs in* $O(n^2 \,\text{polylog}(n))$ *time.*

Using this primitive, we can initially set $\tilde{\mathbf{K}} = 2\gamma(0)\mathbf{I}$ and use it obtain a sparsifier for $\mathbf{K}(0)$ from a linear sketch of $\mathbf{B}$. This sparsifier can then be used on a second sketch of $\mathbf{B}$ to obtain a sparsifier for $\mathbf{K}(1)$, and so on. Working our way up the chain, we can eventually obtain a sparsifier for our original $\mathbf{K}$. While sparsifier recovery will proceed in several levels, we can construct all required sketches in a *single pass* over edge insertions and deletions, and all recovery can be performed in post-processing.

Proof of Theorem 1. Let $\mathbf{K}$ be the Laplacian of our graph G. Process all edge insertions and deletions, using `MaintainSketches` to produce a separate sketch, $(\mathbf{\Pi B})_\ell$ for each $\ell \in \{0, 1, \ldots, \lceil \log_2 \lambda_u/\lambda_l \rceil + 1\}$.

We can use Theorem 3 to recover an ϵ approximation, $\tilde{\mathbf{K}}(\ell)$, for any $\mathbf{K}(\ell)$ given an ϵ approximation for $\mathbf{K}(\ell - 1)$. First, consider the base case, $\mathbf{K}(0)$. Let:

$$\tilde{\mathbf{K}}(0) = \texttt{RefineSparsifier}((\mathbf{\Pi B})_0, \gamma(0)\mathbf{I}, \gamma(0), \epsilon, \frac{1}{2})$$

By Theorem 2, Relation 3:

$$\frac{1}{2}\mathbf{K}(0) \preceq \gamma(0)\mathbf{I} \preceq \mathbf{K}(0)$$

Thus, with high probability, $(1-\epsilon)\mathbf{K}(0) \preceq_r \tilde{\mathbf{K}}(0) \preceq_r (1+\epsilon)\mathbf{K}(0)$ and $\tilde{\mathbf{K}}(0)$ contains $O((1/2)^{-1} \cdot n \log n \cdot \epsilon^{-2}) = O(\epsilon^{-2} n \log n)$ entries.

Now, consider the inductive case. Suppose we have some $\tilde{\mathbf{K}}(\ell-1)$ such that $(1-\epsilon)\mathbf{K}(\ell-1) \preceq_r \tilde{\mathbf{K}}(\ell-1) \preceq_r (1+\epsilon)\mathbf{K}(\ell-1)$. Let:

$$\tilde{\mathbf{K}}(\ell) = \texttt{RefineSparsifier}((\mathbf{\Pi B})_\ell, \frac{1}{2(1+\epsilon)}\tilde{\mathbf{K}}(\ell-1), \gamma(\ell), \epsilon, \frac{1-\epsilon}{2(1+\epsilon)})$$

By Theorem 2, Relation 2:

$$\frac{1}{2}\mathbf{K}(\ell) \preceq \frac{1}{2}\mathbf{K}(\ell-1) \preceq \mathbf{K}(\ell)$$

Furthermore, by assumption we have the inequalities:

$$\frac{1-\epsilon}{1+\epsilon}\mathbf{K}(\ell-1) \preceq_r \frac{1}{1+\epsilon}\tilde{\mathbf{K}}(\ell-1) \preceq_r \mathbf{K}(\ell-1)$$

Thus:

$$\frac{1-\epsilon}{2(1+\epsilon)}\mathbf{K}(\ell) \preceq_r \frac{1}{2(1+\epsilon)}\tilde{\mathbf{K}}(\ell-1) \preceq_r \mathbf{K}(\ell)$$

So, with high probability `RefineSparsifier` returns $\tilde{\mathbf{K}}(\ell)$ such that $(1-\epsilon)\mathbf{K}(\ell) \preceq_r \tilde{\mathbf{K}}(\ell) \preceq_r (1+\epsilon)\mathbf{K}(\ell)$ and $\tilde{\mathbf{K}}(\ell)$ contains just $O((\frac{2(1+\epsilon)}{1-\epsilon})^2 \epsilon^{-2} n \log n) = O(\epsilon^{-2} n \log n)$ nonzero elements. It is important to note that there is no "compounding of error" in this process. Every $\tilde{\mathbf{K}}(\ell)$ is an ϵ approximation for $\mathbf{K}(\ell)$. Error from using $\tilde{\mathbf{K}}(\ell-1)$ instead of $\mathbf{K}(\ell-1)$ is absorbed by a constant factor increase in the number of rows sampled from $\mathbf{B}$. The corresponding increase in sparsity for $\mathbf{K}(\ell)$ does not compound - in fact Theorem 3 is completely agnostic to the sparsity of the coarse approximation $\tilde{\mathbf{K}}$ used.

Finally, to obtain a bonafide spectral graph sparsifier (a weighted subgraph of our streamed graph), let:

$$\tilde{\mathbf{K}} = \texttt{RefineSparsifier}((\mathbf{\Pi B})_{d+1}, \frac{1}{2(1+\epsilon)}\tilde{\mathbf{K}}(d), 0, \epsilon, \frac{1-\epsilon}{2(1+\epsilon)})$$

As in the inductive case,

$$\frac{1-\epsilon}{2(1+\epsilon)}\mathbf{K} \preceq_r \frac{1}{2(1+\epsilon)}\tilde{\mathbf{K}}(d) \preceq_r \mathbf{K}$$

Thus, it follows that, with high probability, $\tilde{\mathbf{K}}$ has sparsity $O(\epsilon^{-2}n\log n)$ and $(1-\epsilon)\mathbf{K} \preceq_r \tilde{\mathbf{K}} \preceq_r (1+\epsilon)\mathbf{K}$. Since we set γ to 0 for this final step, $\tilde{\mathbf{K}}$ simply equals $\tilde{\mathbf{B}}^\top\tilde{\mathbf{B}}$ for some $\tilde{\mathbf{B}}$ that contains reweighted rows of $\mathbf{B}$. Any vector in the kernel of $\mathbf{B}$ is in the kernel of $\tilde{\mathbf{B}}$, and thus any vector in the kernel of $\mathbf{K}$ is in the kernel of $\tilde{\mathbf{K}}$. Thus, we can strengthen our approximation to:

$$(1-\epsilon)\mathbf{K} \preceq \tilde{\mathbf{K}} \preceq (1+\epsilon)\mathbf{K}$$

We conclude that $\tilde{\mathbf{K}}$ is the Laplacian of some graph H containing $O(\epsilon^{-2}n\log n)$ rescaled edges of G and approximating G spectrally to precision ϵ. Finally, note that we only required $d+1 = O(\log n)$ recovery steps, each running in $O(n^2\,\mathrm{polylog}(n))$ time. Thus, the complete recovery time is $O(n^2\,\mathrm{polylog}(n))$. □

5 Streaming Row Sampling

In this section, we develop the sparsifier refinement subroutine required for the proof of Theorem 1 in Section 4.

Proof of Theorem 3. Outside of the streaming model, given full access to $\mathbf{B}$ rather than just a sketch $\mathbf{\Pi B}$ it is easy to implement `RefineSparsifier` via leverage score sampling. Letting $\oplus$ denote appending the rows of one matrix to another, we can define $\mathbf{B}_\gamma = \mathbf{B} \oplus \sqrt{\gamma(\ell)}\cdot\mathbf{I}$, so $\mathbf{K} = \mathbf{B}^\top\mathbf{B} + \gamma\mathbf{I} = \mathbf{B}_\gamma^\top\mathbf{B}_\gamma$. Since $\tau_i = \mathbf{b}_i^\top\mathbf{K}^+\mathbf{b}_i$ and $c\mathbf{K} \preceq_r \tilde{\mathbf{K}} \preceq_r \mathbf{K}$, for any row of $\mathbf{B}_\gamma$ we have

$$\tau_i \leq \mathbf{b}_i^\top\tilde{\mathbf{K}}^+\mathbf{b}_i \leq \frac{1}{c}\tau_i.$$

Let $\tilde{\tau}_i = \mathbf{b}_i^\top\tilde{\mathbf{K}}^+\mathbf{b}_i$ be the leverage score of $\mathbf{b}_i$ approximated using $\tilde{\mathbf{K}}$. Let $\tilde{\boldsymbol{\tau}}$ be the vector of approximate leverage scores, with the leverage scores of the n rows corresponding to $\sqrt{\gamma(\ell)}\cdot\mathbf{I}$ rounded up to 1. This will include the rows of the identity with probability 1 in each independent sampling. While not strictly necessary, doing so simplifies our analysis in the streaming setting. Using this $\tilde{\boldsymbol{\tau}}$ in Lemma 1, we can obtain $\tilde{\mathbf{K}}_\epsilon \approx_\epsilon \mathbf{K}$ with high probability. Since $\|\tilde{\boldsymbol{\tau}}\|_1 \leq \frac{1}{c}\|\boldsymbol{\tau}\|_1 + n \leq \frac{1}{c}\cdot\mathrm{rank}(\mathbf{B}) + n \leq \frac{n+1}{c}$, we can write $\tilde{\mathbf{K}}_\epsilon = \tilde{\mathbf{B}}_\epsilon^\top\tilde{\mathbf{B}}_\epsilon + \gamma\mathbf{I}$, where $\tilde{\mathbf{B}}_\epsilon$ contains $O(\epsilon^{-2}c^{-1}n\log n)$ reweighted rows of $\mathbf{B}$ with high probability.

The challenge in the semi-streaming setting is actually performing the independent edge sampling given only a sketch of $\mathbf{B}$. The general idea is explained in our overview Section 3, with detailed pseudocode included below. We show that each required computation is possible in the dynamic semi-streaming model, and then prove the correctness of the sampling procedure.

Streaming Sparsifier Refinement

`MaintainSketches`$(\mathbf{B}, \epsilon)$**:**

1. For $j \in 1, 2, ...c_0 \frac{1}{\epsilon^2} \log n$
   (a) For $s \in \{1, ...O(\log n)\}$ let $h_s : E \to \{0, 1\}$ be a uniform hash function. Let $\mathbf{B}_s$ be $\mathbf{B}$ with all rows except those with $\prod_{j \leq s} h_j(e) = 0$ zeroed out. So $\mathbf{B}_s$ is $\mathbf{B}$ with rows sampled independently at rate $\frac{1}{2^s}$. $\mathbf{B}_0$ is simply $\mathbf{B}$.
   (b) Maintain $O(\log n)$ sketchs $\mathbf{\Pi}_0\mathbf{B}_0, \mathbf{\Pi}_1\mathbf{B}_1, ..., \mathbf{\Pi}_{O(\log n)}\mathbf{B}_{O(\log n)}$ where $\{\mathbf{\Pi}_0, \mathbf{\Pi}_1, ...\mathbf{\Pi}_{O(\log n)}\}$ are drawn from the distribution from Lemma 2 with $\eta = \frac{1}{4c_1 \log n}$.

`RefineSparsifier`$(\mathbf{\Pi B}, \tilde{\mathbf{K}}, \gamma, \epsilon, c)$**:**

1. For $j \in 1, 2, ...c_0 \frac{1}{\epsilon^2} \log n$
   (a) Compute $\mathbf{\Pi}_s\mathbf{B}_s\tilde{\mathbf{K}}^+$ for each $s \in \{0, 1, 2, ...O(\log n)\}$.
   (b) For each possible edge e:
      i. Compute $\tilde{\tau}_e = \mathbf{b}_e^\top \tilde{\mathbf{K}}^+ \mathbf{b}_e$. Choose s such that $\min\{1, \tilde{\tau}_e\} \leq \frac{1}{2^s} \leq 2 \cdot \min\{1, \tilde{\tau}_e\}$.
      ii. Compute the vector $\mathbf{\Pi}_s\mathbf{x}_e = \mathbf{\Pi}_s\mathbf{B}_s\tilde{\mathbf{K}}^+\mathbf{b}_e$, and perform the heavy hitters algorithm of Lemma 2, recovering with high probability elements with a $\geq \frac{1}{c_1 \log n}$ fraction of the ℓ_2 weight of $\mathbf{x}_e$, and throwing out any recovered elements with a $< \frac{1}{2c_1 \log n}$ fraction of the weight.
      iii. If $\mathbf{x}_e(e)$ is recovered set $\mathbf{W}_j(e, e) = 2^s$
2. Set $\bar{\mathbf{W}} = \frac{1}{c_0 \epsilon^{-2} \log n} \sum_j \mathbf{W}_j$ and output $\tilde{\mathbf{K}}_\epsilon = \mathbf{B}^\top \bar{\mathbf{W}} \mathbf{B} + \gamma \mathbf{I}$.

Implementation Details in the Semi-Streaming Model.

Note that the iterations of main loop of `MaintainSketches` can be done simultaneously in a single pass over the data stream. The sketches $\mathbf{\Pi}_0\mathbf{B}_0, \ldots, \mathbf{\Pi}_{O(\log n)}\mathbf{B}_{O(\log n)}$ can be stacked and the entire $O(n \operatorname{polylog}(n))$ sized compression is output as $\mathbf{\Pi B}$.

`MaintainSketches` requires $O(n \operatorname{polylog}(n))$ space in total, and can be implemented in the dynamic streaming model. When an edge insertion comes in, use $\{h_s\}$ to compute which $\mathbf{B}_s$'s should contain the inserted edge, and update the corresponding sketches. An edge deletion can be performed simply by updating the sketches to reflect adding $-\mathbf{b}_e$ to $\mathbf{B}_s$.

Step 1(a) of `RefineSparsifier` can also be implemented in $O(n \operatorname{polylog} n)$ space. Since $\tilde{\mathbf{K}}$ has $O(n \operatorname{polylog} n)$ nonzeros and since each $\mathbf{\Pi}_s\mathbf{B}_s$ has $O(\operatorname{polylog} n)$ rows, this step simply requires solving $O(\operatorname{polylog} n)$ linear systems on $\tilde{\mathbf{K}}$, which can be performed in $O(n \operatorname{polylog} n)$ time by using a nearly linear time SDD system solver [KMP11]. From this time bound we immediately know that this computation can be performed in $O(n \operatorname{polylog} n)$ space.

In step 1(b)i. it is always possible to choose an appropriate s with $\min\{1, \tilde{\tau}_e\} \leq \frac{1}{2^s} \leq 2 \cdot$

$\min\{1, \tilde{\tau}_e\}$. $\lambda_{max}(\mathbf{K}) \leq n + \gamma = O(\text{poly}(n))$. So $\lambda_{min}(\tilde{\mathbf{K}}^+) = \Omega(\text{poly}(n))$ so $\tilde{\tau}_e = \Omega(\text{poly}(n))$ for all e. So such an s always can be found if we have $O(\log n)$ samplings of $\mathbf{B}$.

Finally, with high probability, when running Step 1(b) for each edge, in total we only ever recover $O(n \log n)$ edges and so can store them in small space.

Correctness

We need to show that, with high probability, in each round of sampling, this algorithm independently samples each row of $\mathbf{B}$ with probability $\hat{\tau}_e$ where $\min\{1, \tilde{\tau}_e\} \leq \hat{\tau}_e \leq 2 \cdot \min\{1, \tilde{\tau}_e\}$. Given this fact, since the algorithm samples the n rows of $\sqrt{\gamma} \cdot \mathbf{I}$ with probability 1, and since $\tau_e \leq \min\{1, \tilde{\tau}_e\} \leq \frac{1}{c}\tau_e$ for all e, by Lemma 1, with high probability, $\tilde{\mathbf{K}}_\epsilon \approx_\epsilon \mathbf{K}$ and $\tilde{\mathbf{K}}_\epsilon = \mathbf{B}\bar{\mathbf{W}}\mathbf{B} + \gamma\mathbf{I} = \tilde{\mathbf{B}}_\epsilon^\top \tilde{\mathbf{B}}_\epsilon + \gamma\mathbf{I}$, where $\tilde{\mathbf{B}}_\epsilon$ contains $O(\epsilon^{-2}c^{-1}n \log n)$ reweighted rows of $\mathbf{B}$.

In the above algorithm, an edge is only included in $\tilde{\mathbf{K}}_\epsilon$ if it is included in the sampled matrix $\mathbf{B}_{s(e)}$ where

$$\min\{1, \tilde{\tau}_e\} \leq \frac{1}{2^{s(e)}} \leq 2 \cdot \min\{1, \tilde{\tau}_e\}$$

The probability of $\mathbf{b}_e$ being included in $\mathbf{B}_{s(e)}$ is simply $1/2^{s(e)}$, and sampling is done independently using uniform random hash functions. So, as long as we can show that with high probability, all $\mathbf{b}_e$ are recovered by the sparse recovery procedure if included in their respective $\mathbf{B}_{s(e)}$, then we are done.

Let $\mathbf{x}_e = \mathbf{B}\tilde{\mathbf{K}}^+\mathbf{b}_e$ and $\mathbf{x}_e^{s(e)} = \mathbf{B}_{s(e)}\tilde{\mathbf{K}}^+\mathbf{b}_e$. If e is not an edge in the original graph or $\mathbf{b}_e$ is not included in $\mathbf{B}_{s(e)}$ then $\mathbf{x}_e^{s(e)}(e) = 0$, so if index e is recovered, it will be discarded. We need to argue that, if $\mathbf{b_e}$ is in fact included in $\mathbf{B}_{s(e)}$, with high probability, $\|\mathbf{x}_e^{s(e)}\|^2$ is not too large, so we are able to identify $\mathbf{x}_e^{s(e)}(e)$. We have:

$$\mathbf{x}_e^{s(e)}(e) = \mathbf{x}_e(e) = \mathbf{1}_e\mathbf{B}\tilde{\mathbf{K}}^+\mathbf{b}_e = \mathbf{b}_e^\top\tilde{\mathbf{K}}^+\mathbf{b}_e = \tilde{\tau}_e \tag{1}$$

Further, we can compute:

$$\begin{aligned}
\|\mathbf{x}_e\|^2 &= \mathbf{b}_e^\top\tilde{\mathbf{K}}^+\mathbf{B}^\top\mathbf{B}\tilde{\mathbf{K}}^+\mathbf{b}_e \\
&\leq \mathbf{b}_e^\top\tilde{\mathbf{K}}^+\mathbf{B}_\gamma^\top\mathbf{B}_\gamma\tilde{\mathbf{K}}^+\mathbf{b}_e && \text{(Since } \mathbf{B}^\top\mathbf{B} \preceq \mathbf{B}_\gamma^\top\mathbf{B}_\gamma\text{)} \\
&\leq \frac{1}{c} \cdot \mathbf{b}_e^\top\tilde{\mathbf{K}}^+\mathbf{b}_e && \text{(Since } c\left(\mathbf{B}_\gamma^\top\mathbf{B}_\gamma\right) \preceq \tilde{\mathbf{K}}\text{)} \\
&\leq \frac{1}{c}\tilde{\tau}_e
\end{aligned}$$

For any edge $e' \neq e$ we define:

$$\tilde{\tau}_{e',e} \stackrel{\text{def}}{=} \mathbf{x}_e^{s(e)}(e') = \mathbf{1}_{e'}\mathbf{B}\tilde{\mathbf{K}}^+\mathbf{b}_e = \mathbf{b}_{e'}^\top\tilde{\mathbf{K}}^+\mathbf{b}_e$$

Lemma 3. $\tilde{\tau}_{e',e} \leq \tilde{\tau}_e$

Proof. Consider $\tilde{\mathbf{v}}_e = \tilde{\mathbf{K}}^+\mathbf{b}_e$. When $\mathbf{K}^+$ is a graph Laplacian $\tilde{\mathbf{v}}_e$ can be interpreted as the approximate voltages induced over each vertex when we treat our edges as resistors and route one

unit of current between the endpoints of e. Letting $e = (u_1, u_2)$ and $e' = (u_1', u_2')$, if we have $|\tilde{\mathbf{v}}_e(u_1') - \tilde{\mathbf{v}}_e(u_2)'| \leq |\tilde{\mathbf{v}}_e(u_1) - \tilde{\mathbf{v}}_e(u_2)|$ then:

$$\mathbf{b}_{e'}^\top \tilde{\mathbf{v}}_e = \mathbf{b}_{e'}^\top \tilde{\mathbf{K}}^+ \mathbf{b}_e \leq \mathbf{b}_e^\top \tilde{\mathbf{K}}^+ \mathbf{b}_e = \mathbf{b}_e^\top \tilde{\mathbf{v}}_e$$

So:

$$\tilde{\tau}_{e',e} \leq \tilde{\tau}_e$$

Now, $\tilde{\mathbf{K}}$ is a weighted graph Laplacian added to a weighted identity matrix. So it is full rank and diagonally dominant. So $\tilde{\mathbf{K}}\tilde{\mathbf{v}}_e = \tilde{\mathbf{K}}\tilde{\mathbf{K}}^+\mathbf{b}_e = \mathbf{b}_e$

Since $\tilde{\mathbf{K}}$ is diagonally dominant and since $\mathbf{b}_e$ is zero everywhere except at $\mathbf{b}_e(u_1) = 1$ and $\mathbf{b}_e(u_2) = -1$, it must be that $\tilde{\mathbf{v}}_e(u_1)$ is the maximum value of $\tilde{\mathbf{v}}_e$ and $\tilde{\mathbf{v}}_e(u_2)$ is the minimum value. So $|\tilde{\mathbf{v}}_e(u_1') - \tilde{\mathbf{v}}_e(u_2)'| \leq |\tilde{\mathbf{v}}_e(u_1) - \tilde{\mathbf{v}}_e(u_2)|$ and $\tilde{\tau}_{e',e} \leq \tilde{\tau}_e$.

□

Now we upper bound the probability that in step (b)ii of `RefineSparsifier` we can't recover edge e from $\mathbf{B}_{s(e)}$ given that it is included in the sample.

$$\begin{aligned}\mathbb{P}\left(\frac{\mathbf{x}_e^{s(e)}(e)^2}{\|\mathbf{x}_e^{s(e)}\|^2} < \frac{1}{c_1 \cdot \log n} \middle| e \in \mathbf{B}_{s(e)}\right) &= \mathbb{P}\left(\|\mathbf{x}_e^{s(e)}\|^2 > c_1 \log n \cdot \tilde{\tau}_e^2\right) \\ &= \mathbb{P}\left(\|\frac{1}{\tilde{\tau}_e}\mathbf{x}_e^{s(e)}\|^2 > c_1 \log n\right)\end{aligned}$$

Note that the vector $\frac{1}{\tilde{\tau}_e}\mathbf{x}_e^{s(e)}$ has all entries (and thus all squared entries) in $[0, 1]$ (by Lemma 3) so we can apply a Chernoff bound to show concentration for its norm. Specifically, we will use a common multiplicative bound [BG11]:

$$\mathbb{P}(X > (1+\delta)\,\mathbb{E}\,X) < e^{-\frac{\delta^2}{2+\delta}\,\mathbb{E}\,X} \tag{2}$$

Recall that

$$\mathbb{E}\,\|\frac{1}{\tilde{\tau}_e}\mathbf{x}_e^{s(e)}\|^2 = \frac{1}{2^{s(e)}} \cdot \frac{\tilde{\tau}_e}{c} \cdot \frac{1}{\tilde{\tau}_e^2} \leq \frac{2}{c} = \Theta(1) \tag{3}$$

which gives

$$\mathbb{P}\left(\|\frac{1}{\tilde{\tau}_e}\mathbf{x}_e^{s(e)}\|^2 > c_1 \log n\right) \leq \mathbb{P}\left(\|\frac{1}{\tilde{\tau}_e}\mathbf{x}_e^{s(e)}\|^2 > \frac{c_1 c \log n}{2}\,\mathbb{E}\,\|\frac{1}{\tilde{\tau}_e}\mathbf{x}_e^{s(e)}\|^2\right) = O(n^{-\Theta(1)})$$

since $\delta = \Theta(\log n)$.

Recall that c is the constant determined by our input coarse sparsifier and c_1 can be chosen by implementing our sparse recovery routine with a different parameter. If we set c_1 large enough, as long as edge e is included in $\mathbf{B}_{s(e)}$, it is recovered with high probability. This guarantee holds for all $\binom{n}{2}$ possible edges with high probability by a union bound. So with high probability, our sampling process is exactly equivalent to independently sampling each edge with probability $\frac{1}{2^{s(e)}}$ where $\min\{1, \tilde{\tau}_e\} \leq \frac{1}{2^{s(e)}} \leq 2 \cdot \min\{1, \tilde{\tau}_e\}$. So our algorithm returns the desired $\tilde{\mathbf{K}}_\epsilon$ with high probability.

□

6 Sparsification of Weighted Graphs

We can use a standard technique to extend our result to streams of weighted graphs in which an edge's weight is specified at deletion, matching what is known for cut sparsifiers in the dynamic streaming model [AGM12b, GKP12]. Assume that all edge weights and the desired approximation factor ϵ are polynomial in n, then we can consider the binary representation of each edge's weight, out to $O(\log n)$ bits. For each bit of precision, we maintain a separate unweighted graph $G_0, G_1, ...G_{O(\log n)}$. We add each edge to the graphs corresponding to bits with value one in its binary representation. When an edge is deleted, its weight is specified, so we can delete it from these same graphs. We have that: G $= \sum_i 2^i \cdot G_i$, so given a $(1 \pm \epsilon)$ sparsifier $\tilde{\mathbf{K}}_i$ for each $\mathbf{K}_i$ we have:

$$(1-\epsilon)\sum_i 2^i \cdot \mathbf{K}_i \preceq \sum_i 2^i \cdot \tilde{\mathbf{K}}_i \preceq (1+\epsilon)\sum_i 2^i \cdot \mathbf{K}_i$$

$$(1-\epsilon)\mathbf{K} \preceq \sum_i 2^i \cdot \tilde{\mathbf{K}}_i \preceq (1+\epsilon)\mathbf{K}$$

So $\sum_i 2^i \cdot \tilde{\mathbf{K}}_i$ is a spectral sparsifier for $\mathbf{K}$, the Laplacian of the weighted graph G.

7 Sparsification of Structured Matrices

Here we show that our algorithm can be extended to handle certain general matrices rather than just graph Laplacians. There were only three places in our analysis where we used that $\mathbf{B}$ was not an arbitrary matrix. First, we needed that $\mathbf{B} = \mathbf{S}\mathbf{B}_n$, where $\mathbf{B}_n$ is the vertex edge incidence matrix of the unweighted complete graph on n vertices. In other words, we assumed that we had some dictionary matrix $\mathbf{B}_n$ whose rows encompass every possible row that could arrive in the data stream. In addition to this dictionary assumption, we needed $\mathbf{B}$ to be sparse and to have a bounded condition number in order to achieve our small space results. These conditions allow our compression to avoid an $\Omega(n^2 \operatorname{polylog}(n))$ lower bound for approximately solving regression on general $\mathbb{R}^{m\times n}$ matrices in the streaming model [CW09].

As such, to handle the general 'structured matrix' case, we assume that we have some dictionary $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{m\times n}$ containing rows $\mathbf{a}_i \in \mathbb{R}^n$ for each $i \in [m]$. We assume that $m = O(\operatorname{poly}(n))$. In the dynamic streaming model we receive insertions and deletions of rows from $\boldsymbol{\mathcal{A}}$ resulting in a matrix $\mathbf{A} = \mathbf{S}\boldsymbol{\mathcal{A}}$ where $\mathbf{S} \in \mathbb{R}^{m\times m}$ is a diagonal matrix such that $\mathbf{S}_{ii} \in \{0, 1\}$ for all $i \in [m]$. Our goal is to recover an $O(n \operatorname{polylog}(m))$ space compression a diagonal matrix $\mathbf{W}$ with at most $O(n \log(n))$ nonzero entries such that $\boldsymbol{\mathcal{A}}^\top \mathbf{W}^2 \boldsymbol{\mathcal{A}} \approx_\epsilon \boldsymbol{\mathcal{A}}^T \mathbf{S}^2 \boldsymbol{\mathcal{A}} = \mathbf{A}^\top \mathbf{A}$. Formally, we prove the following:

Theorem 4 (Streaming Structured Matrix Sparsification)**.** *Given a row dictionary $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{m\times n}$ containing all possible rows of the matrix $\mathbf{A}$, there exists an algorithm that, for any $\epsilon > 0$, processes a stream of row insertions and deletions for $\mathbf{A}$ in a single pass and maintains a set of linear sketches of this input in $O\left(\frac{1}{\epsilon^2} n \operatorname{polylog}(m, \kappa_u)\right)$ space where κ_u is an upper bound on the condition number of $\mathbf{A}^\top \mathbf{A}$. From these sketches, it is possible to recover, with high probability, a matrix $\tilde{\mathbf{A}}^\top \tilde{\mathbf{A}}$ such that $\tilde{\mathbf{A}}$ contains only $O(\epsilon^{-2} n \log n)$ reweighted rows of $\mathbf{A}$ and $\tilde{\mathbf{A}}^T \tilde{\mathbf{A}}$ is a $(1 \pm \epsilon)$ spectral sparsifier of $\mathbf{A}^T \mathbf{A}$. The algorithm recovers $\tilde{A}$ in $\operatorname{poly}(m, \epsilon, n, \log \kappa_u)$ time.*

Note that, when $\mathbf{A}, \kappa_u = o(\operatorname{poly}(n))$, the sketch space is $O\left(\frac{1}{\epsilon^2} n \operatorname{polylog}(n)\right)$. To prove Theorem 4, we need to introduce a more complicated sampling procedure than what was used for the graph

case. In Lemma 3, for the correctness proof of `RefineSparsifier` in Section 5, we relied on the structure of our graph Laplacian and vertex edge incidence matrix to show that $\tilde{\tau}_{e',e} \leq \tilde{\tau}_e$. This allowed us to show that the norm of a sampled $\mathbf{x}_e^{s(e)}$ concentrates around its mean. Thus, we could recover edge e with high probability if it was in fact included in the sampling $\mathbf{B}_{s(e)}$. Unfortunately, when processing general matrices, $\tilde{\tau}_e$ is not necessarily the largest element $\mathbf{x}_e^{s(e)}$ and the concentration argument falls apart.

We overcome this problem by modifying our algorithm to compute more sketches. Rather than computing a single $\mathbf{\Pi A}_s$, for every sampling rate $1/2^s$, we compute $O(\log n)$ sketches of different samplings of $\mathbf{A}$ at rate $1/2^s$. Each sampling is fully independent from the *all* others, including those at the same and different rates. This differs from the graph case, where $\mathbf{B}_{1/2^{s+1}}$ was always a subsampling of $\mathbf{B}_{1/2^s}$ (for ease of exposition). Our modified set up lets us show that, with high probability, the norm of $\mathbf{x_i}^{s(i)}$ is close to its expectation for at least a $(1-\epsilon)$ fraction of the independent samplings for rate $s(i)$. We can recover row i if it is present in one of the 'good' samplings.

Ultimately, we argue, in a similar manner to [KP12], that we can sample rows according to some distribution that is close to the distribution obtained by independently sampling rows according to leverage score. Using this primitive, we can proceed as in the previous sections to prove Theorem 4. In Section 7.1, we provide the row sampling subroutine and in Section 7.2, we show how to use this sampling routine to prove Theorem 4.

7.1 Generalized Row Sampling

Here we show how to sample rows from $\mathbf{A}$ with probability proportional to their leverage scores in the streaming model.

Streaming Row Sampling Algorithm

`MaintainMatrixSketches`$(\mathbf{A}, \epsilon, \kappa_u, \gamma, c)$**:**

1. Let $S = O(\log \kappa_u), T = O(\log m)$ and for all $s \in [S]$ and $t \in [T]$ let $\mathbf{F}_s^{(t)} \in \mathbb{R}^{m \times m}$ be a diagonal matrix such that $[\mathbf{F}_s^{(t)}]_{ii} = 1$ independently with probability $\frac{1}{2^s}$ and is 0 otherwise.[2]
2. For all $s \in [S]$ and $t \in [T]$ maintain sketch $\mathbf{\Pi}_s^{(t)}\mathbf{F}_s^{(t)}\mathbf{A}$ where each $\mathbf{\Pi}_s^{(t)}$ is drawn independently from the distribution in Lemma 2 with $\eta = \frac{1}{4C}$ and $C = O(c^{-1}\epsilon^{-1}\log m)$.
3. Add the rows of $\gamma\mathbf{I}$, sampled appropriately, to the sketches

`RowSampleMatrix`$(\mathbf{\Pi A}, \tilde{\mathbf{K}}, \epsilon, c)$**:**

1. For all $s \in [S]$ and $t \in [T]$ let $\mathbf{x}_s^{(t)} = \mathbf{F}_s^{(t)}\mathbf{A}\tilde{\mathbf{K}}^+$ and compute $\mathbf{\Pi}_s^{(t)}\mathbf{x}_s^{(t)}$.
2. For every $i \in [m]$:
   (a) Compute $\tilde{\tau}_i = \mathbf{a}_i^\top \tilde{\mathbf{K}}^+ \mathbf{a}_i$ and s_i such that $\min\{1, \tilde{\tau}_i\} \leq \frac{1}{2^{s_i}} < 2\min\{1, \tilde{\tau}_i\}$.

(b) Pick $t_i \in [T]$ uniformly at random and use Lemma 2 to check if $\mathbf{x}_{s_i}^{(t_i)}(i)^2 \geq C^{-1}\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2$.

(c) If i is recovered, add row i to the set of sampled edges with weight 2^{s_i}.

We claim that, with high probability, the set of edges returned by the above algorithm is a random variable that is stochastically dominated by the two random variables obtained by sampling edges independently at rates $\tilde{\tau}_i$ and $(1-\epsilon)\tilde{\tau}_i$, respectively..

The following property of PSD matrices is used in our proof of correctness:

Lemma 4. *For any symmetric PSD matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ and indices $i, j \in [n]$ we have*

$$|\mathbf{K}_{ij}| \leq \frac{1}{2}\left(\mathbf{K}_{ii} + \mathbf{K}_{jj}\right)$$

Proof. Let $\mathbf{1}_i$ be the vector with a 1 at position i and 0s else where. For all $i, j \in [n]$ by the fact that $\mathbf{K}$ is PSD we have that

$$(\mathbf{1}_i - \mathbf{1}_j)\,\mathbf{K}\,(\mathbf{1}_i - \mathbf{1}_j) \geq 0 \quad \text{and} \quad (\mathbf{1}_i + \mathbf{1}_j)\,\mathbf{K}\,(\mathbf{1}_i + \mathbf{1}_j) \geq 0$$

Expanding, we have that:

$$-\mathbf{K}_{ii} - \mathbf{K}_{jj} \leq 2\mathbf{K}_{ij} \leq \mathbf{K}_{ii} + \mathbf{K}_{jj}$$

yielding the result. □

We can now proceed to prove that our sampling procedure approximates sampling the rows of $\mathbf{A}$ by their leverage scores.

Lemma 5. *Consider an execution of* `RowSampleMatrix`$(\mathbf{\Pi A}, \tilde{\mathbf{K}}, c, \epsilon)$ *where*

- $c\mathbf{A}^\top\mathbf{A} \preceq \tilde{\mathbf{K}} \preceq \mathbf{A}^\top\mathbf{A}$ *for* $c \in (0, 1]$, *and*
- $\epsilon \in (0, 1]$.

Let $\mathcal{D}$ be a random variable for the indices returned by `RowSampleMatrix`$(\mathbf{\Pi A}, \tilde{\mathbf{K}}, c, \epsilon)$. *Let $\mathcal{I} \subseteq [m]$ denote the indices of the nonzero rows of $\mathbf{A}$ and let $\mathcal{D}_p$ and $\mathcal{D}_q$ be random variables for the subset of $[m]$ obtained by including each $i \in \mathcal{I}$ independently with probability*

$$p_i = (1-\epsilon)\frac{1}{2^{s_i}} \text{ and } q_i = \frac{1}{2^{s_i}}.$$

With high probability, i.e. except for a $(1 - \frac{1}{m^{O(1)}})$ fraction of the probability space, $\mathcal{D}$ is stochastically dominated by $\mathcal{D}_q$ and $\mathcal{D}$ stochastically dominates $\mathcal{D}_p$ with respect to set inclusion.

Proof. By definition, $\mathcal{D}_p$ and $\mathcal{D}_q$ are always subsets of $\mathcal{I}$ and $\mathcal{D}$ is a subset of $\mathcal{I}$ with high probability (it is a subset as long as the algorithm of Lemma 2 succeeds.). Thus it remains to show that, with high probability, for each $\mathcal{J} \subseteq \mathcal{I}$ we have:

$$\prod_{i\in\mathcal{J}} p_i = \mathbb{P}[\mathcal{J} \subseteq \mathcal{D}_p] \leq \mathbb{P}[\mathcal{J} \subseteq \mathcal{D}] \leq \mathbb{P}[\mathcal{J} \subseteq \mathcal{D}_q] = \prod_{i\in\mathcal{J}} q_i.$$

[2]Throughout this section, for $X \in \mathbb{Z}^+$ we let $[X] = \{0, 1, 2, \ldots, X\}$

Furthermore, by definition, with high probability, `RowSampleMatrix` outputs $i \in \mathcal{I}$ if and only if $\mathbf{x}_{s_i}^{(t_i)}(i)^2 \geq C^{-1}\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2$ and consequently

$$\mathbb{P}[\mathcal{J} \subseteq \mathcal{D}] = \mathbb{P}\left[\forall i \in \mathcal{J} \ : \ \mathbf{x}_{s_i}^{(t_i)}(i)^2 \geq C^{-1}\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2\right]. \tag{4}$$

As shown in Equation 1, when proving our graph sampling Lemma, for all $i \in \mathcal{J}$,

$$\mathbf{x}_{s_i}^{(t_i)}(i) = [\mathbf{F}_{s_i}^{(t_i)}]_{ii} \cdot \tilde{\tau}_i.$$

Consequently, by the definition of $[\mathbf{F}_{s_i}^{(t_i)}]_{ii}$ we can rewrite (4) as:

$$\mathbb{P}[\mathcal{J} \subseteq \mathcal{D}] = \mathbb{P}\left[\forall i \in \mathcal{J} \ : \ \|\mathbf{x}_{s_i}^{(t_i)}\|_2^2 \leq C \cdot \tilde{\tau}_i^2 \text{ and } [\mathbf{F}_{s_i}^{(t_i)}]_{ii} = 1\right]. \tag{5}$$

From (5) and the independence of $[\mathbf{F}_{s_i}^{(t_i)}]_{ii}$ we obtain the following trivial upper bound on $\mathbb{P}[\mathcal{J} \subseteq \mathcal{D}]$,

$$\mathbb{P}[\mathcal{J} \subseteq \mathcal{D}] \leq \mathbb{P}\left[\forall i \in \mathcal{J} \ : \ [\mathbf{F}_{s_i}^{(t_i)}]_{ii} = 1\right] = \prod_{i \in \mathcal{J}} \frac{1}{2^{s_i}} = \prod_{i \in \mathcal{J}} q_i$$

and consequently $\mathcal{D}$ is stochastically dominated by $\mathcal{D}_q$ as desired.

As shown in Equation **??**, when proving the graph sampling case, for all $i \in \mathcal{I}$ and $t \in [T]$

$$\mathbb{E}\left[\|\mathbf{x}_{s_i}^{(t)}\|_2^2\right] \leq \frac{2}{c}\tilde{\tau}_i^2. \tag{6}$$

Combining (5) and (6) yields

$$\mathbb{P}[\mathcal{J} \subseteq \mathcal{D}] \geq \mathbb{P}\left[\forall i \in \mathcal{J} \ : \ \|\mathbf{x}_{s_i}^{(t_i)}\|_2^2 \leq \frac{c}{2} \cdot C \cdot \mathbb{E}[\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2] \text{ and } [\mathbf{F}_{s_i}^{(t_i)}]_{ii} = 1\right]. \tag{7}$$

To bound the probability that $\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2 \leq \frac{c}{2} \cdot C \cdot \mathbb{E}[\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2]$ we break the contribution to $\|\mathbf{x}_{s_i}^{(t)}\|_2^2$ for each t into two parts. For all i we let $\mathcal{K}_i = \{j \in \mathcal{I} | s_j = s_i\}$, i.e. the set of all rows j which we attempt to recover at the same sampling rate as i. For any $t \in [T]$, we let $A_i^{(t)} = \sum_{j \in \mathcal{K}} x_i^{(t)}(j)^2$ and $B_i^{(t)} = \sum_{j \in \mathcal{I}-\mathcal{K}} x_i^{(t)}(j)^2$. Using this notation and (7) we obtain the following lower bound

$$\mathbb{P}[\mathcal{J} \subseteq \mathcal{D}] \geq \mathbb{P}\left[\forall i \in \mathcal{J} \ : \ A_i^{(t_i)} \leq \frac{C \cdot c}{4} \cdot \mathbb{E}[\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2] \ , \ B_i^{(t_i)} \leq \frac{C \cdot c}{4} \cdot \mathbb{E}[\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2] \text{ , and } [\mathbf{F}_{s_i}^{(t_i)}]_{ii} = 1\right].$$

Furthermore since $\mathbf{K} \preceq \mathbf{A}^\top \mathbf{A}$, for i such that $s_i \geq 1$ we have

$$\mathbb{E}\left[\|\mathbf{x}_{s_i}^{(t)}\|_2^2\right] \geq \frac{1}{2^{s_i}} \cdot \mathbf{a}_i^\top \tilde{\mathbf{K}}^+ \mathbf{A}^\top \mathbf{A} \tilde{\mathbf{K}}^+ \mathbf{a}_i \geq \tilde{\tau}_i^2 \ . \tag{8}$$

For all $j \in \mathcal{K}_i$, the rows that we attempt to recover at the same rate as row i, we know that $\tilde{\tau}_j \leq 2\tilde{\tau}_i$. By Lemma 4 we know that for all $i \in \mathcal{I}$ with $s_i \geq 1$ and $j \in \mathcal{K}_i$

$$x_{s_i}^{(t)}(j)^2 = [\mathbf{F}_{s_i}^{(t)}]_{jj} \cdot \left|\mathbf{a}_i^\top \tilde{\mathbf{K}}^+ \mathbf{a}_j\right|^2 \leq 1 \cdot \left(\frac{\tilde{\tau}_i + \tilde{\tau}_j}{2}\right)^2 \leq \left(\frac{\tilde{\tau}_i + 2\tilde{\tau}_i}{2}\right)^2 \leq 3\,\mathbb{E}[\|\mathbf{x}_{s_i}^{(t)}\|_2^2] \ . \tag{9}$$

Now recall that $C = O(c^{-1}\epsilon^{-1}\log m)$ and $T = O(\log m)$. If $\tilde{\tau}_i > 1/2$ and therefore $s_i = 0$ then $|\mathbf{x}_{s_i}^{(t_i)}(i)| = \tilde{\tau}_i$ and considering (6) we see that row i is output with high probability. On the

other hand if $s_i \geq 1$, then by (9) and Chernoff bound choosing a sufficiently large constant in $C = O(c^{-1}\epsilon^{-2}\log m)$ we can ensure that with high probability $A_i^{(t)} \leq \frac{C \cdot c}{4} \cdot \mathbb{E}[\|\mathbf{x}_{s_i}^{(t)}\|_2^2]$ for all i and t.

Furthermore, by (6) and Markov bound we know that $\mathbb{P}[B_i^{(t_i)} > \frac{C \cdot c}{4}] \leq \frac{1}{O(\epsilon^{-1}\log m)}$. Therefore, by Chernoff bound, with high probability for each $i \in \mathcal{J}$ with $s_i \geq 1$ for at least a $1-\epsilon$ fraction of the values of $t \in T$ we have $B_i^{(t_i)} \leq \frac{C \cdot c}{4} \cdot \mathbb{E}[\|\mathbf{x}_{s_i}^{(t_i)}\|_2^2]$. However, note that by construction all the $B_i^{(t)}$ are mutually independent of the $A_i^{(t)}$ and the values of $[\mathbf{F}_{s_j}^{(t)}]_{jj}$ for $j \in K_i$. So, `RowSampleMatrix` is simply picking each row i with probability $\frac{1}{2^{s_i}}$ (failing with only a $\frac{1}{m^{O(1)}}$ probability) or not being able to recover each edge independently with some probability at most ϵ - the probability that $B_i^{(t_i)}$ is too large. Consequently, except for a negligible fraction of the probability space we have that

$$\mathbb{P}\left[\mathcal{J} \subseteq \mathcal{D}\right] \geq \prod_{i \in \mathcal{J}} (1-\epsilon) \cdot [\mathbf{F}_{s_i}^t]_{ii} = \prod_{i \in \mathcal{J}} \frac{1-\epsilon}{2^{s_i}} = \prod_{i \in \mathcal{J}} p_i$$

and we have the desired result. □

7.2 Generalized Recursive Sparsification

Here, we show how to construct a spectral sparsifier in the streaming model for a general structured matrix using the row sampling subroutine, `RowSampleMatrix`. In the graph case, Theorem 1 shows that, if we can find a sparsifier to a graph G using a coarse sparsifier, then we can use the chain of spectrally similar graphs provided in Theorem 2 to find a final $(1 \pm \epsilon)$ sparsifier for our input graph.

The proof of Theorem 1 includes our third reliance on the fact that we are sparsifying graphs - we show that the condition number of an unweighted graph is polynomial in n. This fact does not hold in the general matrix case - the condition number can be exponentially large even for bounded integer matrices. Therefore, our result for general matrix depends on the condition number of $\mathbf{A}$. We now show the key ingredient to proving 4, the general matrix analog to `RefineSparsifier` (Theorem 3).

Theorem 5. *Given a row dictionary $\boldsymbol{\mathcal{A}} \in \mathbb{R}^{m \times n}$. Let $\mathbf{A} = \mathbf{S}\boldsymbol{\mathcal{A}}$ be the matrix specified by an insertion-deletion stream where $\mathbf{S} \in \mathbb{R}^{m \times m}$ is a diagonal matrix such that $\mathbf{S}_{ii} \in \{0, 1\}$ for all $i \in [m]$. Let κ_u be a given upper bound on the possible condition number of any $\mathbf{A}$. Let γ be a fixed parameter and consider $\mathbf{K} = \mathbf{A}^\top\mathbf{A} + \gamma\mathbf{I}$. For any $\epsilon > 0$, there exists a sketching procedure `MaintainMatrixSketches`$(\mathbf{A}, \epsilon, \gamma, \kappa_u, c)$ that outputs an $O(n \operatorname{polylog}(m, \kappa_u))$ sized sketch $\boldsymbol{\Pi}\mathbf{A}$. There exists a corresponding recovery algorithm `RefineMatrixSparsifier` such that if $c\mathbf{K} \preceq \tilde{\mathbf{K}} \preceq \mathbf{K}$ for some $0 < c < 1$ then:*

`RefineMatrixSparsifier`$(\boldsymbol{\Pi}\mathbf{A}, \tilde{\mathbf{K}}, \epsilon, c)$ returns, with high probability, $\tilde{\mathbf{K}}_\epsilon = \tilde{\mathbf{A}}_\epsilon^\top \tilde{\mathbf{A}}_\epsilon + \gamma\mathbf{I}$, where $(1-\epsilon)\mathbf{K} \preceq_r \tilde{\mathbf{K}}_\epsilon \preceq_r (1+\epsilon)\mathbf{K}$, and $\tilde{\mathbf{A}}_\epsilon$ contains only $O(\epsilon^{-2} n \log n)$ reweighted rows of $\mathbf{A}$ with high probability.

Proof. As in the graph case, we can think of the identity $\gamma\mathbf{I}$ as a set of rows that we sample with probability 1. Hence, we have $\tilde{\mathbf{K}}_\epsilon = \tilde{\mathbf{A}}_\epsilon^\top \tilde{\mathbf{A}}_\epsilon + \gamma\mathbf{I}$.

Lemma 5 shows that `RowSampleMatrix`$(\boldsymbol{\Pi}\mathbf{A}, \tilde{\mathbf{K}}, c, \epsilon)$ returns a random set of indices of $\boldsymbol{\mathcal{A}}$ such that the generated random variable is dominated by $\mathcal{D}_q$ and is stochastically dominates $\mathcal{D}_p$. Recall

that $\mathcal{D}_p$ and $\mathcal{D}_q$ are random variables for the subset of $[m]$ obtained by including each $i \in \mathcal{I}$ independently with probability

$$p_i = (1-\epsilon)\frac{1}{2^{s_i}} \text{ and } q_i = \frac{1}{2^{s_i}}.$$

Since $\frac{1}{2^{s_i}}$ is a constant factor approximation of leverages score, Lemma 1 shows that sampling and reweighing the rows according to $\mathcal{D}_p$ for $O(\epsilon^{-2}\log n)$ trials gives a spectral sparsifier of $\mathbf{K}$ with the guarantee required. Similarly, sampling according to $\mathcal{D}_q$ gives a sparsifier. Since the indices returned by $\texttt{RowSampleMatrix}(\mathbf{A}, \tilde{\mathbf{K}}, c, \epsilon)$ are sandwiched between two processes which each give spectral sparsifiers, sampling according to `RowSampleMatrix` for $O\left(\epsilon^{-2}\log n\right)$ trials gives the required spectral sparsifier [KP12]. □

Using `RefineMatrixSparsifier`, the arguments in Theorem 1 yield Theorem 4. Our sketch size needs to be based on $\log \kappa_u$ for two reasons - we must subsample the matrix at $O(\log \kappa_u)$ different rates as our leverage scores will be lower bounded by some $\text{poly}(\kappa_u)$. Further the chain of recursive sparsifiers presented in Theorem 2 will have length $\log \kappa_u$. Recovery will run in time $\text{poly}(m, n, \epsilon, \log \kappa_u)$. Its space usage will depend on the sparsity of the rows in $\mathbf{A}$ as we will need enough space to solve linear systems in $\tilde{\mathbf{K}}$. In the worst case, this will require $O(n^2)$ space, however, if the row of $\mathbf{A}$ are sparse, and hence $\tilde{\mathbf{K}}$ is sparse, recovery will take less space - $O(n\,\text{polylog}(m))$ with constant row sparsity.

8 Using a Pseudorandom Number Generator

In the proof of our sketching algorithm, Theorem 3, we assume that `MaintainSketches` has access to $O(\log n)$ uniform random hash functions, $h_1, \ldots, h_{O(\log n)}$ mapping every edge to $\{0, 1\}$. These functions are used to subsample our vertex edge incidence matrix, $\mathbf{B}$, at geometrically decreasing rates. Storing the functions as described would require $O(n^2 \log n)$ space - we need $O(\log n)$ random bits for each possible edge.

To achieve $O(n\,\text{polylog}(n))$ space, we need to compress the hash functions using Nisan's pseudorandom number generator. Our approach follows an argument in [AGM12b] (Section 3.4) that was originally introduced in [Ind00] (Section 3.3). First, we summarize the pseudorandom number generator from [Nis92]

Theorem 6 (Corollary 1 in [Nis92])**.** *Any randomized algorithm running in space(S) and using R random bits may be converted to one that uses only $O(S \log R)$ random bits (and runs in space $(O(S \log R))$)*

[Nis92] gives this conversion explicitly by describing a method for generating R pseudorandom bits from $O(S \log R)$ truly random bits. For any algorithm running in *space*(S), the pseudorandom bits are "good enough" in that the probability of reaching any possible end state of the algorithm will not change noticeably. Such a bound is possible because every reachable end state in a *space*(S) algorithm must show up with probability at *minimum* 2^{-S}. By starting with $O(S \log R)$ truly random bits, the pseudorandom generator achieves an additive error of 2^{-S} on end state probabilities. Thus, the failure probability of any randomized algorithm at most doubles.

Now, Theorem 6 does not apply to our algorithm as is. Our issue is not the *use* of $O(n^2 \log n)$ truly random bits, but rather the fact we have to keep all of the bits in memory for the entire

duration of `MaintainSketches`. We might receive an update to any of the $O(n^2)$ possible edges at any time, so the corresponding hash function needs to be available. We cannot simply say that we will use a pseudorandom generator to expand $O(\log(n^2 \log n))$ random bits whenever the hash functions are needed, thus giving us a small space algorithm that can handle the pseudorandomess - that argument is circular.

Instead, consider the following: suppose our algorithm is used on a sorted edge stream where all insertions and deletions for a single edge come in consecutively. In this case, at any given time, we only need to store one random bit for each hash function, which requires just $O(\log n)$ space. The random bits can be discarded after moving on to the next edge. Thus, the entire algorithm could truly run in $O(n \operatorname{polylog}(n))$ space. Then, we can apply Theorem 6, using the pseudorandom generator to get all of our required random bits by expanding just $O(\log(n^2 \log n)) = O(\log n)$ bits and still succeeding with high probability. We will no longer have to discard the random bits to run in $O(n \operatorname{polylog} n)$ space. Now notice that, since our algorithm is sketch based, edge updates simply requires an addition to or subtraction from a sketch matrix. These operations commute, so our output will not differ if we reorder of the insertion/deletion stream. Thus, we can run our algorithm on a general edge stream, using the pseudorandom number generator to generate any of the required $O(n^2 \log n)$ bits as they are needed and operating in only $O(n \operatorname{polylog} n)$ space.

Each time an edge is streamed in, we need to generate $\log n$ random bits from the pseudorandom generator. This can be done in $\log(R) * S = O(n \operatorname{polylog}(n))$ time [Ind00], which dominates the runtime required to process each streaming update.

Finally, Section 7 uses a slightly different sampling scheme for general structured matrices. Instead of building a sequence of subsampled matrices, the row dictionary is sampled independently at each level. In total, the required number of random bits is $O(m \log^2 n)$, where m is the number of rows in the dictionary $\mathbf{A}$. We require that $m = \operatorname{poly}(n)$, in which case the arguments above apply unmodified for the general matrix case.

9 Acknowledgements

We would like to thank Richard Peng for pointing us to the recursive row sampling algorithm contained in [MP12], which became a critical component of our streaming algorithm. We would also like to thank Jonathan Kelner for useful discussions and Jelani Nelson for a helpful initial conversation on oblivious graph compression.

This work was partially supported by NSF awards 0843915, 1111109, and 0835652, CCF-1065125, CCF-AF-0937274, CCF-0939370, and CCF-1217506, NSF Graduate Research Fellowship grant 1122374, Hong Kong RGC grant 2150701, AFOSR grants FA9550-13-1-0042 and FA9550-12-1-0411, MADALGO center, Simons Foundation, and the Defense Advanced Research Projects Agency (DARPA).

References

[AG09] Kook Jin Ahn and Sudipto Guha. Graph sparsification in the semi-streaming model. In *ICALP (2)*, pages 328–338, 2009.

[AGM12a] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Analyzing graph structure via linear measurements. In *SODA*, pages 459–467, 2012.

[AGM12b] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Graph sketches: sparsification, spanners, and subgraphs. In *PODS*, pages 5–14, 2012.

[AGM13] Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Spectral sparsification in dynamic graph streams. In *APPROX-RANDOM*, pages 1–10, 2013.

[BG11] Avrim Blum and Anupam Gupta. Randomized algorithms lecture notes. http://www.cs.cmu.edu/~avrim/Randalgs11/lectures/lect0124.pdf, 2011.

[BK96] András Benczúr and David Karger. Approximating s-t minimum cuts in $\tilde{O}(n^2)$ time. In *STOC*, pages 47–55, 1996.

[CW09] Kenneth Clarkson and David Woodruff. Numerical linear algebra in the streaming model. In *STOC*, pages 205–214, 2009.

[Elk11] Michael Elkin. Streaming and fully dynamic centralized algorithms for constructing and maintaining sparse spanners. *ACM Transactions on Algorithms*, 7(2):20, 2011.

[ELMS11] Leah Epstein, Asaf Levin, Julián Mestre, and Danny Segev. Improved approximation guarantees for weighted matching in the semi-streaming model. *SIAM Journal on Discrete Mathematics*, 25(3):1251–1265, 2011.

[FKM+05] Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. On graph problems in a semi-streaming model. *Theoretical Computer Science*, 348(2):207–216, 2005.

[GI10] Anna C. Gilbert and Piotr Indyk. Sparse recovery using sparse matrices. *Proceedings of the IEEE*, 98(6):937–947, 2010.

[GKP12] Ashish Goel, Michael Kapralov, and Ian Post. Single pass sparsification in the streaming model with edge deletions. *CoRR*, abs/1203.4900, 2012.

[GLPS10] Anna Gilbert, Yi Li, Ely Porat, and Martin Strauss. Approximate sparse recovery: optimizing time and measurements. *STOC*, pages 475–484, 2010.

[Har12] Nick Harvey. Matrix concentration. http://www.cs.rpi.edu/~drinep/RandNLA/slides/Harvey_RandNLA@FOCS_2012.pdf, 2012.

[HRR99] Monika R. Henzinger, Prabhakar Raghavan, and Sridhar Rajagopalan. External memory algorithms. chapter Computing on Data Streams, pages 107–118. American Mathematical Society, Boston, MA, USA, 1999.

[Ind00] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings and data stream computation. In *FOCS*, pages 189–197, 2000.

[KL11] Jonathan A Kelner and Alex Levin. Spectral sparsification in the semi-streaming setting. In *STACS*, page 440, 2011.

[KMP11] Ioannis Koutis, Gary L. Miller, and Richard Peng. A nearly-m log n time solver for sdd linear systems. In *FOCS*, pages 590–598, 2011.

[KP12] Michael Kapralov and Rina Panigrahy. Spectral sparsification via random spanners. *ITCS*, pages 393–398, 2012.

[KW14] Michael Kapralov and David Woodruff. Spanners and sparsifiers in dynamic streams. *manuscript*, pages 107–118, 2014.

[LMP13] Mu Li, Gary L. Miller, and Richard Peng. Iterative row sampling. In *FOCS*, pages 127–136, 2013.

[McG13] Andrew McGregor. Graph stream algorithms: A survey. http://people.cs.umass.edu/~mcgregor/papers/13-graphsurvey.pdf, 2013.

[MP12] Gary L. Miller and Richard Peng. Iterative approaches to row sampling. *CoRR*, abs/1211.2713v1, 2012.

[Mut05] S. Muthukrishnan. Data streams: Algorithms and applications. www.cs.rutgers.edu/~muthu/stream-1-1.ps, 2005.

[Nis92] Noam Nisan. Pseudorandom generators for space-bounded computation. *Combinatorica*, 12(4):449–461, 1992.

[NN13] Jelani Nelson and Huy L. Nguyen. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *FOCS*, pages 117–126, 2013.

[Sar06] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *FOCS*, pages 143–152, 2006.

[SS08] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *STOC*, pages 563–568, 2008.

[ST04] Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *STOC*, pages 81–90, 2004.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

A Spectral Sparsifiers Via Leverage Score Sampling

Here we give the proof of Lemma 1, the primitive used to obtain a spectral sparsifier via leverage score sampling.

Lemma (Spectral Sparsifier via Leverage Score Sampling)**.** *Let $\tilde{\boldsymbol{\tau}}$ be a vector of m estimated leverage scores for the rows of $\mathbf{B}$, such that $1 \geq \tilde{\tau}_i \geq \tau_i$ for all $i \in [m]$. For some known constant c, let $\mathbf{W_1}, \mathbf{W_2}, \ldots \mathbf{W}_{c \log n \epsilon^{-2}}$ be diagonal matrices with independently chosen entries such that $\mathbf{W_j}(i,i) = \frac{1}{\tilde{\tau}_i}$ with probability $\tilde{\tau}_i$ and $\mathbf{W_j}(i,i) = 0$ otherwise. Letting $\bar{\mathbf{W}} = \frac{1}{c \log n \epsilon^{-2}} \cdot \sum_j \mathbf{W_j}$ then with high probability,*

$$\tilde{\mathbf{K}} = \mathbf{B}^\top \bar{\mathbf{W}} \mathbf{B} \approx_\epsilon \mathbf{K}$$

Furthermore, $\bar{\mathbf{W}}$ *has* $O(\|\tilde{\boldsymbol{\tau}}\|_1 \log n\epsilon^{-2})$ *nonzeros with high probability. That is, if we sample each each row of* $\mathbf{B}$ *independently with probability* $\tilde{\tau}_i$*, reweight selected rows by* $\frac{1}{\sqrt{\tilde{\tau}_i}}$*, and average over* $c \log n\epsilon^{-2}$ *trials, we obtain a matrix* $\tilde{\mathbf{B}} = \mathbf{B}\bar{\mathbf{W}}^{1/2}$ *such that* $\tilde{\mathbf{B}}^\top \tilde{\mathbf{B}} = \mathbf{B}^\top \bar{\mathbf{W}} \mathbf{B} = \tilde{\mathbf{K}} \approx_\epsilon \mathbf{K}$ *and* $\tilde{\mathbf{B}}$ *contains just* $O(\|\tilde{\boldsymbol{\tau}}\|_1 \log n\epsilon^{-2})$ *reweighted rows of* $\mathbf{B}$ *with high probability.*

In our proof we use a variant of Corollary 5.2 of [Tro12], given by Harvey in [Har12]

Lemma 6. *Let* $\mathbf{Y_1} ... \mathbf{Y_k}$ *be independent random positive semidefinite matrices of size* $n \times n$*. Let* $\mathbf{Y} = \sum_i \mathbf{Y_i}$ *and let* $\mathbf{Z} = \mathbb{E}[\mathbf{Y}]$*. If* $\mathbf{Y_i} \preceq R \cdot \mathbf{Z}$ *then*

$$\mathbb{P}\left[\sum_i \mathbf{Y_i} \preceq (1-\epsilon)\mathbf{Z}\right] \leq n e^{\frac{-\epsilon^2}{2R}}$$

and

$$\mathbb{P}\left[\sum_i \mathbf{Y_i} \succeq (1+\epsilon)\mathbf{Z}\right] \leq n e^{\frac{-\epsilon^2}{3R}}$$

This matrix concentration result gives us the foundation for our leverage score sampling Lemma.

Proof of Lemma 1. For each row $\mathbf{b_i}$ of $\mathbf{B}$ let $\mathbf{Y_i} = \frac{\mathbf{b_i}\mathbf{b_i}^\top}{\tilde{\tau}_i \cdot c \log n\epsilon^{-2}}$ with probability $\tilde{\tau}_i$, and 0 otherwise. We have:

$$\mathbf{Y_i} \preceq \frac{\mathbf{b_i}\mathbf{b_i}^\top}{\tilde{\tau}_i \cdot c \log n\epsilon^{-2}} \preceq \frac{\mathbf{b_i}\mathbf{b_i}^\top}{\tau_i \cdot c \log n\epsilon^{-2}} \preceq \frac{1}{c \log n\epsilon^{-2}} \cdot \mathbf{K} \tag{10}$$

since

$$\frac{\mathbf{b_i}\mathbf{b_i}^\top}{\tau_i} = \frac{\mathbf{b_i}\mathbf{b_i}^\top}{\mathbf{b_i}^\top \mathbf{K}^+ \mathbf{b_i}} \preceq \mathbf{K} \tag{11}$$

This can be proven by showing that for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}^\top \mathbf{b_i}\mathbf{b_i}^\top \mathbf{x} \leq \mathbf{b_i}^\top \mathbf{K}^+ \mathbf{b_i} \cdot \mathbf{x}^\top \mathbf{K} \mathbf{x}$. We can assume without loss of generality that $\mathbf{x}$ is in the column space of $\mathbf{K}$, since letting $\mathbf{x}'$ be the component of $\mathbf{x}$ in the null space of $\mathbf{K}$, $\mathbf{x'}^\top \mathbf{K} \mathbf{x'} = \mathbf{x'}^\top \mathbf{b_i}\mathbf{b_i}^\top \mathbf{x'} = 0$. This means that for some $\mathbf{y}$ we can write $\mathbf{x} = \mathbf{K}^{+/2}\mathbf{y}$ where $\mathbf{K}^{+/2} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{V}^\top$ (recalling that $\mathbf{K}^+ = \mathbf{V}\boldsymbol{\Sigma}^{-2}\mathbf{V}^\top$). So now we consider:

$$\mathbf{y}^\top (\mathbf{K}^{+/2} \mathbf{b_i}\mathbf{b_i}^\top \mathbf{K}^{+/2}) \mathbf{y}$$

By the cyclic property of trace we know that $\mathrm{tr}(\mathbf{K}^{+/2} \mathbf{b_i}\mathbf{b_i}^\top \mathbf{K}^{+/2}) = \mathrm{tr}(\mathbf{b_i}^\top \mathbf{K}^+ \mathbf{b_i}) = \tau_i$. And since it is a PSD matrix, its maximum eigenvalue is at most τ_i. So

$$\mathbf{y}^\top (\mathbf{K}^{+/2} \mathbf{b_i}\mathbf{b_i}^\top \mathbf{K}^{+/2}) \mathbf{y} \leq \tau_i \|\mathbf{y}\|^2 = \tau_i \cdot \mathbf{y}^\top \mathbf{K}^{+/2} \mathbf{K} \mathbf{K}^{+/2} \mathbf{y}$$

which gives us Equation 11

Now, since $\mathbb{E}[\mathbf{Y_i}] = \frac{1}{c \log n\epsilon^{-2}} \cdot \mathbf{b_i}\mathbf{b_i}^\top$, If we take $c \log n\epsilon^{-2}$ independent samples of each $\mathbf{Y_i}$ and add them together, our expected sum is $\mathbf{K}$. So by Lemma 6 and Equation 10, we have:

$$\tilde{\mathbf{K}} = \mathbf{B}^\top \bar{\mathbf{W}} \mathbf{B} \approx_\epsilon \mathbf{K}$$

with probability:

$$\geq 1 - n e^{\frac{-c \log n\epsilon^{-2} \epsilon^2}{3}} \leq 1 - n^{1-c/3} \tag{12}$$

Further, for each i, $\bar{\mathbf{W}}(i,i) \geq 0$ with probability $\geq \tilde{\tau}_i$ and $\leq c \log n\epsilon^{-2} \tilde{\tau}_i$. Since $\sum_i \tilde{\tau}_i \geq \sum_i \tau_i \geq n - 1$, by a Chernoff bound, $\bar{\mathbf{W}}$ has $O(\|\tilde{\boldsymbol{\tau}}\|_1 \log n\epsilon^{-2})$ nonzero entries with high probability. □

B Sparse Recovery

In this section we give a proof of the ℓ_2 heavy hitters algorithm given in Lemma 2. It is known that ℓ_2 heavy hitters is equivalent to the ℓ_2/ℓ_2 sparse recovery problem [GI10]. Some sparse recovery algorithms are in fact based on algorithms for solving heavy hitters problem. However, we were not able to find a suitable reference for an ℓ_2 heavy hitters algorithm so we show the reduction here - namely, how to find ℓ_2 heavy hitters using a sparse recovery algorithm.

We follow the terminology of [GLPS10]. An approximate sparse recovery system consists of parameters k, N, an $m \times N$ measurement matrix $\mathbf{\Phi}$, and a decoding algorithm D. For any vector $\mathbf{x} \in \mathbb{R}^N$ the decoding algorithm D can be used to recover an approximation $\widehat{\mathbf{x}}$ to $\mathbf{x}$ from the *linear sketch* $\mathbf{\Phi x}$. In this paper we will use a sparse recovery algorithm that achieves the ℓ_2/ℓ_2 sparse recovery guarantee:

$$||\widehat{\mathbf{x}} - \mathbf{x}||_2 \leq C \cdot ||\mathbf{x} - \mathbf{x}_k||_2$$

where $\mathbf{x}_k$ is the best k-term approximation to $\mathbf{x}$ and $C > 1$. Our main sparse recovery primitive is the following result of [GLPS10]:

Theorem 7 (Theorem 1 in [GLPS10])**.** *For each $k \geq 1$ and $\epsilon > 0$, there is an algorithm and a distribution $\mathbf{\Phi}$ over matrices in $\mathbb{R}^{O(k \log(N/k)/\epsilon) \times N}$ satisfying that for any $\mathbf{x} \in \mathbb{R}^N$, given $\mathbf{\Phi x}$, the algorithm returns $\widehat{\mathbf{x}}$ such that $\widehat{\mathbf{x}}$ has $O(k \log^{O(1)} N/\epsilon)$ non-zeros and*

$$||\widehat{\mathbf{x}} - \mathbf{x}||_2^2 \leq (1+\epsilon)||\mathbf{x} - \mathbf{x}_k||_2^2$$

with probability at least $3/4$. The decoding algorithm runs in time $O(k \log^{O(1)} N/\epsilon)$.

Using this primitive, we can prove Lemma B.

Lemma (ℓ_2 Heavy Hitters)**.** *For each $\eta > 0$, there is a decoding algorithm D and a distribution on matrices $\mathbf{\Phi}$ in $\mathbb{R}^{O(\eta^{-2} \operatorname{polylog}(N)) \times N}$ such that, for any $\mathbf{x} \in \mathbb{R}^N$, with probability $1 - N^{-c}$ over the choice of $\mathbf{\Phi}$, given $\mathbf{\Phi x}$, the algorithm D returns a vector $\mathbf{w}$ such that $\mathbf{w}$ has $O(\eta^{-2} \operatorname{polylog}(N))$ non-zeros and satisfies*

$$||\mathbf{x} - \mathbf{w}||_\infty \leq \eta ||\mathbf{x}||_2.$$

with probability $1 - N^{-c}$. The sketch $\mathbf{\Phi x}$ can be maintained and decoded in $O(\eta^{-2} \operatorname{polylog}(N))$ space.

Proof. Let $h : [N] \to [16/\eta^2]$ be a random hash function (pairwise independence suffices), and for $j = 1, \ldots, 16/\eta^2$ let $\mathbf{y}_i^j = \mathbf{x}_i$ if $h(i) = j$ and 0 o.w. For a vector $\mathbf{u} \in \mathbb{R}^N$ we write $\mathbf{u}_{-i}$ to denote $\mathbf{u}$ with the i-th component zeroed out.

By Markov's inequality we have

$$\mathbb{P}[||\mathbf{y}_{-i}^{h(i)}||^2 > \eta^2 \|\mathbf{x}_{-i}\|^2/2] < 1/8.$$

Note that since we are only using Markov's inequality, it is sufficient to have h be pairwise independent. Such a function h can be represented in small space. Now invoke the result of Theorem 7 on $\mathbf{y}^{h(i)}$ with $k = 1$, $\epsilon = 1$, and let $\mathbf{w}^{h(i)}$ be the output. We have

$$||\mathbf{y}^{h(i)} - \mathbf{w}^{h(i)}||_2^2 \leq 2||\mathbf{y}^{h(i)} - \mathbf{y}_k^{h(i)}||_2^2 \leq 2||\mathbf{y}_{-i}^{h(i)}||^2.$$

Hence, we have

$$(\mathbf{y}_i^{h(i)} - \mathbf{w}_i^{h(i)})^2 \leq \eta^2 \|\mathbf{x}\|^2.$$

This shows that applying sketches from Theorem 7 to vectors $\mathbf{y^j}$, for $j = 1, \ldots, 16/\eta^2$ and outputting the vector $\mathbf{w}$ with $\mathbf{w_i} = \mathbf{w}_i^{h(i)}$ allows us to recover all $i \in [N]$ with $\eta\|\mathbf{x}\|_2$ additive error with probability at least $3/4 - 1/8$.

Performing $O(\log N)$ repetitions and taking the median value of $\mathbf{w_i}$ yields the result. Note that our scheme uses $O(\eta^{-2}\,\mathrm{polylog}(N))$ space and decoding time, and is linear in $\mathbf{x}$, as desired. □

C Recursive Sparsification

For completeness, we give a short proof of Theorem 2:

Theorem (Recursive Sparsification ([MP12], Section 4))**.** *Consider any PSD matrix* $\mathbf{K}$ *with maximum eigenvalue bounded from above by* λ_u *and minimum nonzero eigenvalue bounded from below by* λ_l*. Let* $d = \lceil \log_2(\lambda_u/\lambda_l) \rceil$*. For* $\ell \in \{0, 1, 2, ..., d\}$*, define:*

$$\gamma(\ell) = \lambda_u / 2^\ell$$

So, $\gamma(d) \leq \lambda_l$ *and* $\gamma(0) = \lambda_u$*. Then the chain of PSD matrices,* $\left[\mathbf{K}(0), \mathbf{K}(1), \ldots, \mathbf{K}(d)\right]$ *with:*

$$\mathbf{K}(\ell) = \mathbf{K} + \gamma(\ell)\mathbf{I}_{n \times n}$$

satisfies the following relations:

1. $\mathbf{K} \preceq_r \mathbf{K}(d) \preceq_r 2\mathbf{K}$
2. $\mathbf{K}(\ell) \preceq \mathbf{K}(\ell - 1) \preceq 2\mathbf{K}(\ell)$ *for all* $\ell \in \{1, \ldots, d\}$
3. $\mathbf{K}(0) \preceq 2\gamma(0)\mathbf{I} \preceq 2\mathbf{K}(0)$

When $\mathbf{K}$ *is the Laplacian of an unweighted graph,* $\lambda_{max} < 2n$ *and* $\lambda_{min} > 8/n^2$ *(where here* λ_{min} *is the smallest* nonzero *eigenvalue). Thus the length of our chain,* $d = \lceil \log_2 \lambda_u/\lambda_l \rceil$*, is* $O(\log n)$*.*

Proof. Relation 1 follows trivially from the fact that $\gamma(d) \leq \lambda_l$ is smaller than the smallest nonzero eigenvalue of $\mathbf{K}$. For any $\mathbf{x} \perp \ker(\mathbf{K})$:

$$\mathbf{x}^\top \mathbf{K}(d)\mathbf{x} = \mathbf{x}^\top \mathbf{K}\mathbf{x} + \mathbf{x}^\top(\gamma(d)\mathbf{I})\mathbf{x} \leq \mathbf{x}^\top \mathbf{K}\mathbf{x} + \mathbf{x}^\top(\lambda_{min}\mathbf{I})\mathbf{x} \leq 2\mathbf{x}^\top \mathbf{K}\mathbf{x}$$

The other direction follows from $\gamma(d)\mathbf{I} \succeq 0$. Using the same argument, relation 3 follows from the fact that $\gamma(0) \geq \lambda_{max}(\mathbf{K})$. For relation 2:

$$2\mathbf{K}(\ell) = 2\mathbf{K} + 2\gamma(\ell)\mathbf{I} = 2\mathbf{K} + \gamma(\ell - 1)\mathbf{I} \succeq \mathbf{K}(\ell - 1)$$

Again, the other direction just follows from $\gamma(\ell)\mathbf{I} \succeq 0$.

Finally, we need to prove the required eigenvalue bounds. For an unweighted graph, $\lambda_{max} < n$ follows from fact that n is the maximum eigenvalue of the Laplacian of the complete graph on n vertices. $\lambda_{min} > 8/n^2$ by Lemma 6.1 of [ST04]. Note that this argument extends to weighted graphs when the ratio between the heaviest and lightest edge is bounded by a polynomial in n. □