Streaming Complexity of Spanning Tree Computation

Yi-Jun Chang
ETH Zürich, Switzerland
yi-jun.chang@eth-its.ethz.ch

Martín Farach-Colton
Rutgers University, USA
farach@cs.rutgers.edu

Tsan-Sheng Hsu
Academia Sinica, Taipei City, Taiwan
tshsu@iis.sinica.edu.tw

Meng-Tsung Tsai
National Chiao Tung University, Hsinchu, Taiwan
mtsai@cs.nctu.edu.tw

──── **Abstract** ────

The semi-streaming model is a variant of the streaming model frequently used for the computation of graph problems. It allows the edges of an n-node input graph to be read sequentially in p passes using $\tilde{O}(n)$ space. If the list of edges includes deletions, then the model is called the turnstile model; otherwise it is called the insertion-only model. In both models, some graph problems, such as spanning trees, k-connectivity, densest subgraph, degeneracy, cut-sparsifier, and $(\Delta + 1)$-coloring, can be exactly solved or $(1 + \varepsilon)$-approximated in a single pass; while other graph problems, such as triangle detection and unweighted all-pairs shortest paths, are known to require $\tilde{\Omega}(n)$ passes to compute. For many fundamental graph problems, the tractability in these models is open. In this paper, we study the tractability of computing some standard spanning trees, including BFS, DFS, and maximum-leaf spanning trees.

Our results, in both the insertion-only and the turnstile models, are as follows.

Maximum-Leaf Spanning Trees: This problem is known to be APX-complete with inapproximability constant $\rho \in [245/244, 2)$. By constructing an ε-*MLST sparsifier*, we show that for every constant $\varepsilon > 0$, MLST can be approximated in a single pass to within a factor of $1 + \varepsilon$ w.h.p. (albeit in super-polynomial time for $\varepsilon \leq \rho - 1$ assuming P $\neq$ NP) and can be approximated in polynomial time in a single pass to within a factor of $\rho_n + \varepsilon$ w.h.p., where ρ_n is the supremum constant that MLST cannot be approximated to within using polynomial time and $\tilde{O}(n)$ space. In the insertion-only model, these algorithms can be deterministic.

BFS Trees: It is known that BFS trees require $\omega(1)$ passes to compute, but the naïve approach needs $O(n)$ passes. We devise a new randomized algorithm that reduces the pass complexity to $O(\sqrt{n})$, and it offers a smooth tradeoff between pass complexity and space usage. This gives a polynomial separation between single-source and all-pairs shortest paths for unweighted graphs.

DFS Trees: It is unknown whether DFS trees require more than one pass. The current best algorithm by Khan and Mehta [STACS 2019] takes $\tilde{O}(h)$ passes, where h is the height of computed DFS trees. Note that h can be as large as $\Omega(m/n)$ for n-node m-edge graphs. Our contribution is twofold. First, we provide a simple alternative proof of this result, via a new connection to sparse certificates for k-node-connectivity. Second, we present a randomized algorithm that reduces the pass complexity to $O(\sqrt{n})$, and it also offers a smooth tradeoff between pass complexity and space usage.

1   Introduction

Spanning trees are critical components of graph algorithms, from depth-first search trees (DFS) for finding articulation points and bridges [44], computing *st*-numbering [13], chain decomposition [41], and coloring signed graphs [18], to breadth-first search trees (BFS) for finding separators [33], computing sparse certificates of *k*-node-connectivity [8, 12], approximating diameters [10, 40], and characterizing AT-free graphs [5], and to maximum-leaf spanning trees (MLST) for connected dominating sets [35, 42] and connected maximum cuts [26, 21].

In the semi-streaming model, the tractability of spanning tree computation, except arbitrary spanning trees [3, 43, 39], is less studied. The *semi-streaming* model [37, 3] is a variation of streaming model frequently used for the computation of graph problems. It allows the edges of an n-node input graph to be read sequentially in p passes using $\tilde{O}(n)$[1] space. If the list of edges includes deletions, then the model is called the turnstile model; otherwise it is called the insertion-only model. In both models, some graph problems, such as spanning trees [3], k-connectivity [25], densest subgraph [36], degeneracy [15], cut-sparsifier [29], and $(\Delta + 1)$-coloring [4], can be exactly solved or $(1 + \varepsilon)$-approximated in a single pass, while other graph problems, such as triangle detection and unweighted all-pairs shortest paths [7], are known to require $\tilde{\Omega}(n)$ passes to compute. For many fundamental graph problems, e.g., standard spanning trees, the tractability in these models is open. BFS computation is known to require $\omega(1)$ passes [17], but only the naive $O(n)$-pass algorithm is known. It is unknown whether DFS computation requires more than one passes [14, 30], but the current best algorithm needs $\tilde{O}(h)$ passes [30] where h is the height of the computed DFS trees, so $h = O(n)$ for dense graphs. The tractability of maximum-leaf spanning trees (MLST) is unknown even allowing $O(n^2)$ space, since it is APX-complete [34, 20].

Due to the lack of efficient streaming algorithms for spanning tree computation, for some graph problems that are traditionally solved using spanning trees, such as finding articulation points and bridges, people had to look for alternative methods when designing streaming algorithms for these problems [16, 14]. The alternative methods, even if they are based on known results in graph theory, may still involve the design of new streaming algorithms. For the problems mentioned above, the alternative methods use newly-designed sparse connectivity certificates [12, 25] that are easily computable in the semi-streaming

---

[1] We write $\tilde{O}(k)$ to denote $O(k \operatorname{poly} \log n)$ or $O(k/\operatorname{poly} \log n)$ where n is the number of nodes in the input graph. Similarly, $\tilde{\Omega}(k)$ denotes $\Omega(k \operatorname{poly} \log n)$ or $\Omega(k/\operatorname{poly} \log n)$.

model, rather than the classical one due to Nagamochi and Ibaraki [38]. Hence establishing the hardness of spanning tree computation helps to explain the need of the alternative methods.

In this paper, we study the tractability of computing standard spanning trees for connected simple undirected graphs, including BFS trees, DFS trees, and MLST. Unless otherwise stated, our upper bounds work in the turnstile model (and hence also in the insertion-only model), and our lower bounds hold for the insertion-only model (and hence also in the turnstile model). The space upper and lower bounds are in bits. Our results are as follows.

Maximum-Leaf Spanning Trees: We show, by constructing an ε-*MLST sparsifier* (Theorem 6), that for every constant $\varepsilon > 0$, MLST can be approximated in a single pass to within a factor of $1 + \varepsilon$ w.h.p.[2] (albeit in super-polynomial time for $\varepsilon \leq \rho - 1$ since it is APX-complete [34, 20] with inapproximability constant $\rho \in [245/244, 2)$ [9]) and can be approximated in polynomial time in a single pass to within a factor of $\rho_n + \varepsilon$ w.h.p., where ρ_n is the supremum constant that MLST cannot be approximated to within using polynomial time and $\tilde{O}(n)$ space. In the insertion-only model, these algorithms are deterministic. We also show a complementary hardness result (Theorem 17) that for every $k \in [1, (n-5)/4]$, to approximate MLST to within an additive error k, any single-pass randomized streaming algorithm that succeeds with probability at least $2/3$ requires $\Omega(n^2/k^2)$ bits. This hardness result excludes the possibility to have a single-pass semi-streaming algorithm to approximate MLST to within an additive error $n^{1/2-\Omega(1)}$. Our results for MLST shows that intractability in the sequential computation model (i.e., Turing machine) does not imply intractability in the semi-streaming model.

Our algorithms rely on a new sparse certificate, the ε-*MLST sparsifier*, defined as follows. Let G be an n-node m-edge connected simple undirected graph. Then for any given constant $\varepsilon > 0$, H is an ε-*MLST sparsifier* if it is a connected spanning subgraph of G with $|E(H)| \leq f(\varepsilon)|V(G)|$ and $\mathrm{leaf}(H) \geq (1-\varepsilon)\,\mathrm{leaf}(G)$, where $\mathrm{leaf}(G)$ denotes the maximum number of leaves (i.e. nodes of degree one) that any spanning tree of G can have and f is some function independent of n. We show that an ε-MLST sparsifier can be constructed efficiently in the semi-streaming model.

▶ **Theorem 1.** *In the turnstile model, for every constant $\varepsilon > 0$, there exists a randomized algorithm that can find an ε-MLST sparsifier with probability $1 - 1/n^{\Omega(1)}$ using a single pass, $\tilde{O}(f(\varepsilon)n)$ space, and $\tilde{O}(n+m)$ time, and in the insertion-only model a deterministic algorithm that uses a single pass, $\tilde{O}(f(\varepsilon)n)$ space, and $O(n+m)$ time.*

Combining Theorem 1 with any polynomial-time RAM algorithms for MLST that uses $\tilde{O}(n+m)$ space, e.g., [34, 35, 42], we obtain the following result.

▶ **Corollary 2.** *In the turnstile model, for every constant $\varepsilon > 0$, there exists a randomized algorithm that can approximate MLST for any n-node connected simple undirected graph with probability $1 - 1/n^{\Omega(1)}$ to within a factor of $\rho_n + \varepsilon$ using a single pass, $\tilde{O}(f(\varepsilon)n)$ space, and polynomial time, where ρ_n is the supremum constant that MLST cannot be approximated to within using polynomial time and $\tilde{O}(n)$ space, and in the insertion-only model a deterministic algorithm that uses a single pass, $\tilde{O}(f(\varepsilon)n)$ space, and polynomial time.*

Using Corollary 2, we show that approximate connected maximum cut can be computed in a single pass using $\tilde{O}(n)$ space for unweighted regular graphs (Corollary 7).

---

[2] W.h.p. means with probability $1 - 1/n^{\Omega(1)}$.

BFS Trees: It is known that BFS trees require $\omega(1)$ passes to compute [17], but the naive approach needs $O(n)$ passes. We devise a randomized algorithm that reduces the pass complexity to $O(\sqrt{n})$ w.h.p., and give a smooth tradeoff between pass complexity and space usage.

▶ **Theorem 3.** *In the turnstile model, for each $p \in [1, \sqrt{n}]$, there exists a randomized algorithm that can compute a BFS tree for any n-node connected simple undirected graph with probability $1 - 1/n^{\Omega(1)}$ in p passes using $\tilde{O}((n/p)^2)$ space, and in the insertion-only model a deterministic algorithm that uses $\tilde{O}(n^2/p)$ space.*

This gives a polynomial separation between single-source and all-pairs shortest paths for unweighted graphs because any randomized semi-streaming algorithm that computes unweighted all-pairs shortest paths with probability at least $2/3$ requires $\tilde{\Omega}(n)$ passes.

We extend Theorem 3 and obtain that multiple BFS trees, each starting from a unique source node, can be computed more efficiently in pass complexity in a batch than individually (see Theorem 13). We show that this batched BFS has applications to computing a 1.5-approximation of diameters for unweighted graphs (Theorem 15) and a 2-approximation of Steiner trees for unweighted graphs (Corollary 14).

DFS Trees: It is unknown whether DFS trees require more than one passes [14, 30], but the current best algorithm needs $\tilde{O}(h)$ passes due to Khan and Mehta [30], where h is the height of computed DFS trees. We devise a randomized algorithm that has pass complexity $O(\sqrt{n})$ w.h.p., and give a smooth tradeoff between pass complexity and space usage.

▶ **Theorem 4.** *In the turnstile model, for each $p \in [1, \sqrt{n}]$, there exists a randomized algorithm that can compute a DFS tree for any n-node connected simple undirected graph with probability $1 - 1/n^{\Omega(1)}$ in p passes that uses $\tilde{O}(n^3/p^4)$ space, and in the insertion-only model a deterministic algorithm that uses $\tilde{O}(n^2/p^2)$ space.*

For dense graphs, our algorithms improves upon the current best algorithms for DFS due to Khan and Mehta [30] which needs $\Omega(m/n)$ passes for n-node m-edge graphs in the worst case because of the existence of (m/n)-cores, where a k-core is a maximal connected subgraph in which every node has at least k neighboring nodes in the subgraph.

1.1 Technical Overview

Maximum-Leaf Spanning Trees: We construct an ε-MLST sparsifier by a new result that complements Kleitman and West's lower bounds on the maximum number of leaves for graphs with minimum degree $\delta \geq 3$ [31]. The lower bounds are: if a connected simple undirected graph G has minimum degree δ for some sufficiently large δ, then $\mathrm{leaf}(G) \geq (1 - (2.5\ln\delta)/\delta)|V(G)|$ and the leading constant can be larger for $\delta \in \{3, 4\}$. Our complementary result (Lemma 5), without the restriction on the minimum degree, is: any connected simple undirected graph G, except the singleton graph, has

$$\mathrm{leaf}(G) \geq \frac{1}{10}(|V(G)| - \mathrm{inode}(G)), \tag{1}$$

where $\mathrm{inode}(G)$ denotes the number of nodes whose degree is two and whose neighbors both have degree two. Equation (1) implies that, if one can find a connected spanning subgraph H of G so that $|\mathrm{leaf}(G) - \mathrm{leaf}(H)| \leq \varepsilon(V(G) - \mathrm{inode}(G))$, then one gets an (10ε)-MLST sparsifier.

Our sparsification technique is general enough to obtain a $(t + \varepsilon)$-approximation for MLST in a single pass using $\tilde{O}(n)$ space by combining any t-approximation $\tilde{O}(n)$-space RAM algorithm for MLST with our ε-MLST sparsifier. On the other hand, since in linear time one can find an ε-MLST sparsifier of $O(n)$ edges, any t-approximation RAM algorithm for MLST with time complexity $O(f(n, m))$ can be reduced to $O(f(n, n) + n + m)$ if a small sacrifice on approximation ratio is allowed. This reduces the time complexity of RAM algorithms for MLST that need superlinear time on the number of edges, such as the local search approach from $O(m^k n^{k+2})$ for $k \geq 1$ to $O(n^{2k+2})$ and the leafy forest approach from $O((m + n)\alpha(n))$ to $O(m + n\alpha(n))$, both due to Lu and Ravi [34, 35].

BFS Trees: We present a simple deterministic algorithm attaining a smooth tradeoff between pass complexity and space usage. In particular, in the insertion-only semi-streaming model, the algorithm finishes in $O(n/\text{poly} \log n)$ passes. The algorithm is based on an observation that the sum of degrees of nodes in any root-to-leaf path of a BFS tree is bounded by $O(n)$ (Lemma 8).

Our more efficient randomized algorithm (Theorem 3) constructs a BFS tree by combining the results of multiple instances of bounded-radius BFS. To reduce the space usage, the simulation of these bounded-radius BFS are assigned random starting times, and the algorithm only maintains the last three layers of each BFS tree. These ideas are borrowed from results on shortest paths computation in the parallel and the distributed settings [11, 22, 27, 45].

DFS Trees: We present a simple alternative proof of the result of Khan and Mehta [30] that a DFS tree can be constructed in $\lceil h/k \rceil$ passes using $\tilde{O}(nk)$ space, for any given parameter k, where h is the height of the computed DFS tree. The new proof is based on the following connection between the DFS computation and the sparse certificates for k-node-connectivity. We show in Lemma 16 that the first k layers of *any* DFS tree of a such a certificate H can be extended to a DFS tree of the original graph G.

The proof of Theorem 4 is based on the parallel DFS algorithm of Aggarwal and Anderson [2]. In this paper we provide an efficient implementation of their algorithm in the streaming model, also via the sparse certificates for k-node-connectivity, which allows us to reduce the number of passes by batch processing.

We note that in a related work, Ghaffari and Parter [23] showed that the parallel DFS algorithm of Aggarwal and Anderson can be adapted to distributed setting. Specifically, they showed that DFS can be computed in the CONGEST model in $\tilde{O}(\sqrt{Dn} + n^{3/4})$ rounds, where D is the diameter of the graph.

1.2   Paper Organization

In Section 2, we present how to construct an ε-MLST sparsifier and apply it to devise single-pass semi-streaming algorithms to approximate MLST to within a factor of $(1 + \varepsilon)$ for every constant $\varepsilon > 0$. Then, in Section 3, we show how to compute a BFS tree rooted at a given node by an $O(\sqrt{n})$-pass $\tilde{O}(n)$-space algorithm w.h.p. and its applications to computing approximate diameters and approximate Steiner trees. In Section 4, we have a similar result for computing DFS trees; that is, $O(\sqrt{n})$-pass $\tilde{O}(n)$-space algorithm that succeeds w.h.p. Lastly, we prove the claimed single-pass lower bound in Section 5.

2 Maximum-Leaf Spanning Trees

In this section, we will show how to construct an ε-MLST sparsifier in the semi-streaming model; that is, proving Theorem 1. We recall the notions defined in Section 1 before proceeding to the results. By *ignorable node*, we denote a node x whose degree is two and whose neighbors u and v have degree two as well. Note that $u \neq v$ for simple graphs. Let $\mathrm{leaf}(G)$ be the maximum number of leaves (i.e. nodes of degree one) that a spanning tree of G can have. Let $\mathrm{inode}(G)$ denote the number of ignorable nodes in G. Let $\deg_G(x)$ denote the degree of node x in graph G. Let $S_k(G)$ denote any subgraph of G so that $S_k(G)$ contains all nodes in G and every node x in $S_k(G)$ has degree $\deg_{S_k}(x) \geq \min\{\deg_G(x), k\}$. Let $T(G)$ be any spanning tree of a connected graph G.

We begin with a result that complements Kleitman and West's lower bounds on the number of leaves for graphs with minimum degree δ for any $\delta \geq 3$. Our lower bound does not rely on the degree constraint. The constant $1/10$ in Lemma 5 may be improved, but the subsequent lemmata and theorems only require it to be $\Omega(1)$.

▶ **Lemma 5.** *Every connected simple undirected graph G, except the singleton graph, has*

$$\mathrm{leaf}(G) \geq \frac{1}{10}(|V(G)| - \mathrm{inode}(G)).$$

Proof. Our proof is a generalization of the dead leaf argument due to Kleitman and West [31]. Let T be a tree rooted at s with $N(s)$ as leaves for some arbitrary node $s \in G$ initially, where $N(s)$ denotes the neighbors of s, and then grow T iteratively by a node expansion order, defined below. By expanding T at node x, we mean to select a leaf node x of T and add all of x's neighbors in $G \backslash T$, say $y_1, y_2, \ldots, y_d$, and their connecting edges, $(x, y_1), (x, y_2), \ldots, (x, y_d)$, to T. In this way, every node outside T cannot be a neighbor of any non-leaf node in T. We say a leaf node in T is *dead* if it has no neighbor in $G \setminus T$. Let $(\Delta n)_i$ denote the number of non-ignorable nodes in G that joins T while the i-th operation is applied. Let $(\Delta \ell)_i$ denote the change of the number of leaf nodes in T while the i-th operation is applied. Let $(\Delta m)_i$ denote the change of the number of dead leaf nodes in T while the i-th operation is applied. The subscript i may be removed when the context is clear. We need to secure that $\Delta \ell + \Delta m \geq \Delta n/5$ holds for each of the following operations and the initial operation.

Operation 1: If T has a leaf node x that has $d \geq 2$ neighbors outside T, then expand T at x. In this case, $\Delta n \leq d$, $\Delta \ell \geq d - 1$, and $\Delta m \geq 0$.

Operation 2: If every leaf node in T has at most one neighbor outside T and some node $x \notin T$ has $d \geq 2$ neighbors in T, then expand T at one of x's neighbors in T. In this case, $\Delta n \leq 1$, $\Delta \ell = 0$, and $\Delta m = d - 1$.

Operation 3: This operation is used only when the previous two operations do not apply. Let x_0 be some leaf in T that has exactly one neighbor x_1 not in T. For each $i \geq 1$, if x_i is defined and all neighbors of x_i other than x_{i-1} are outside T and x_i has degree two in G, then define x_{i+1} to be the neighbor of x_i other than x_{i-1}. Suppose that x_i for $i \leq k$ are defined and x_{k+1} is not defined, then we expand T at x_i for each $i \leq k$ in order. Though k can be arbitrarily large, $\Delta n \leq 2 + \deg_G(x_k)$. If x_{k+1} is not defined and x_k has $d > 0$ neighbors other than x_{k-1} in T (thus $k \geq 2$ in this case otherwise Operation 2 applies), then we discuss in subcases:

Subcase 1 ($\deg_G(x_k) = 1$): It is impossible to have $\deg_G(x_k) = 1$ for this case.
Subcase 2 ($\deg_G(x_k) = 2$): Then $\Delta \ell = 0$ and $\Delta m = 2$.
Subcase 3 ($\deg_G(x_k) \geq 3$): Then $\Delta \ell = \deg_G(x_k) - d - 2$ and $\Delta m \geq d$.

If x_{k+1} is not defined and x_k has 0 neighbor other than x_{k-1} in T, then $\deg_G(x_k)$ is either 1 or ≥ 3. For $\deg_G(x_k) = 1$, $\Delta\ell = 0$ and $\Delta m = 1$. For $\deg_G(x_k) \geq 3$, $\Delta\ell = \deg_G(x_k) - 2$ and $\Delta m \geq 0$.

It is clear that one can expand T to get a spanning tree of G by a sequence of the above operations. Because all leaves are eventually dead, $\sum \Delta m = \sum \Delta\ell$. Consequently, $2\,\mathrm{leaf}(G) \geq 2\sum \Delta\ell = \sum \Delta\ell + \Delta m \geq (\sum \Delta n)/5 = (V(G) - \mathrm{inode}\,G)/5$, as desired. ◀

Given Lemma 5, our goal is, for every constant $\varepsilon > 0$, find a sparse subgraph H of the input graph G so that:

1. The nodes incident to the edges in $T^* \setminus H$ can be *dominated* by a small set S of at most $\varepsilon(|V(G)| - \mathrm{inode}(G))$ nodes, i.e. either in S or has at least one neighbor node in S using the edges in H, where T^* is any optimal MLST of G.
2. H is connected.

Because of the existence of the small dominating set S, one can obtain a forest F from $T^* \cap H$ by adding some edges in H so that the number of leaves in F is no less than that in T^* by $|S|$ and the number of connected components in F is no more than that in T^* by $|S|$. Since H is connected, one can further obtain a spanning tree T from F by adding at most $|S|$ edges in H, so the number of leaves in T is no less than that in F by $2|S|$. Pick an H associated with a sufficiently small ε, by Equation (1) H is an ε-MLST sparsifier. A formal proof is given below.

▶ **Theorem 6.** *For every integer $k \geq 186$, every connected simple undirected graph G has*

$$\mathrm{leaf}(S_k(G) \cup T(G)) \geq \left(1 - 30\left(\frac{1 + \ln(k+1)}{k+1}\right)\right)\mathrm{leaf}(G).$$

Proof. Let T^* be a spanning tree of G that has $\mathrm{leaf}(G)$ leaves. Let k be some fixed integer at least 3 and let $H = S_k(G) \cup T(G)$. Let $L = \{x \in V(G): x \text{ is incident to some } e \in T^* \setminus H\}$. Note that every node $x \in L$ has $\deg_G(x) > k$, so x and all neighbors of x are not ignorable nodes in G.

First, we show that L can be dominated by a small set S of size at most $\varepsilon(|V(G)| - \mathrm{inode}(G))$ using some edges in H. We obtain S from two parts, S_1 and S_2. S_1 is a random node subset sampled from the non-ignorable nodes in G, in which each node is included in S_1 with probability p independently, for some $p \in (0, 1)$ to be determined later. Thus, $E[|S_1|] = p(|V(G)| - \mathrm{inode}(G))$. Since every node $x \in L$ is adjacent only to the non-ignorable nodes in G, the probability that $x \in L$ is not dominated by any node in S_1 is

$$\Pr[x \text{ is not dominated}] = (1-p)^{1+\deg_H(x)} \leq (1-p)^{k+1}.$$

Let S_2 be the set of nodes in L that are not dominated by any node in S_1 using the edges in H. Thus,

$$E[|S|] = E[|S_1| + |S_2|] \leq \left(p + (1-p)^{k+1}\right)(|V(G)| - \mathrm{inode}(G)).$$

Then, we obtain a forest F from $T^* \cap H$ by adding some edges in H as follows. Initially, $F = T^* \cap H$.

Operation 1: For each $x \in L$, if x is an isolated node in $T^* \cap H$ and $x \notin S$, then add an edge e from x to some node in S to F. Such an edge e must exist because S dominates L.

Operation 2: For each $x \in L$, if x is not an isolated node in $T^* \cap H$ and the connected component that contains x has an empty intersection with S, then add an edge e from x to some node in S to F. Again, such an edge e must exist because S dominates L.

For each leaf $\ell \in T^*$, if $\deg_G(\ell) \leq k$, then ℓ is a leaf in $T^* \cap H$ (also in F unless $\ell \in S$); otherwise $\deg_G(\ell) > k$, if ℓ is not a leaf in $T^* \cap H$, then ℓ must be an isolated node in $T^* \cap H$, and by Operation 1 ℓ is connected to some node in S unless $\ell \in S$. Hence, except those in S, every ℓ is a leaf node in F, so the number of leaves in F is no less than that in T^* by $|S|$. By Operation 2, the number of connected component is at most $|S|$.

Lastly, since H is connected, one can obtain a spanning tree T from F by connecting the components in F by some edges in H. Thus, the number of leaves in T is no less than that in T^* by $3|S|$. To obtain an ε-MLST sparsifier, by Lemma 5 we need:

$$\frac{3|S|}{\frac{1}{10}(|V(G)| - \mathrm{inode}(G))} \leq 30\left(p + (1-p)^{k+1}\right) \leq 30\left(p + e^{-p(k+1)}\right) \leq \varepsilon$$

Setting $p = (\ln(k+1))/(k+1)$ gives the desired bound, and the leading constant is positive for $k \geq 186$. ◀

To find such a subgraph H, fetching a spanning tree of the input graph G and grabbing k edges for each node in G suffices. Thus, we get a single-pass $\tilde{O}(n)$-space algorithm for the insertion-only model. As for the turnstile model, we use $\tilde{O}(k)$ ℓ_0-samplers [28] for each node in G to fetch at least k neighbors of x w.h.p., and fetch a spanning tree by appealing to the single-pass $\tilde{O}(n)$-space algorithm for spanning trees in dynamic streams [3]. This gives a proof of Theorem 1.

Applications.   In [21], Gandhi et al. show a connection between the maximum-leaf spanning trees and connected maximum cut. Their results imply that, for any unweighted regular graph G, the connected maximum cut can be found by the following two steps:

Step 1: Find a spanning tree T whose $\mathrm{leaf}(T) \geq (1/2 - \varepsilon)\,\mathrm{leaf}(G)$ for some constant $\varepsilon > 0$.
Step 2: Randomly partition the leaves in T into two parts L and R so that each leaf is included in L with probability $1/2$ independently.

Then, outputting L and $V(G) \setminus L$ yields an $8 + \varepsilon$-approximation for connected maximum cut. Step 1 is the bottleneck and can be implemented by combining our ε-MLST sparsifier (Theorem 1) with the 2-approximation algorithm for MLST due to Solis-Oba, Bonsma, and Lowski [42]. This gives Corollary 7.

▶ **Corollary 7.** *In the turnstile model, for every constant $\varepsilon > 0$, there exists a randomized algorithm that can approximate the connected maximum cut for n-node unweighted regular graphs to within a factor of $8 + \varepsilon$ with probability $1 - 1/n^{\Omega(1)}$ in a single pass using $\tilde{O}(f(\varepsilon)n)$ space.*

3   Breadth-First Search Trees

A BFS tree of an n-node connected simple undirected graph can be constructed in $O(n)$ passes using $\tilde{O}(n)$ space by simulating the standard BFS algorithm layer by layer. By storing the entire graph, a BFS tree can be computed in a single pass using $O(n^2)$ space. In Section 3.1, we show that it is possible to have a smooth tradeoff between pass complexity and space usage. In Section 3.2, we prove Theorem 3, which shows that the above tradeoff can be

improved when randomness is allowed, even in the turnstile model. Then, in Section 3.3, we show that multiple BFS trees, each starting from a distinct source node, can be computed more efficiently in a batch than individually. Lastly, we demonstrate an application to diameter approximation in Section 3.4.

In the BFS problem, we are given an n-node connected simple undirected graph $G = (V, E)$ and a distinguished node s, and it suffices to compute the distance $\text{dist}(s, v)$ for each node $v \in V \setminus \{s\}$. To infer a BFS tree from the distance information $\{\text{dist}(s, v) : v \in V\}$, it suffices to assign a parent to each node $v \in V \setminus \{s\}$ the smallest-identifier node from the set $\{u \in N(v) : \text{dist}(s, u) = \text{dist}(s, v) - 1\}$ where $N(v)$ is the set of v's neighbors. This can be done with one additional pass using $\tilde{O}(n)$ space in the insertion-only model. In the turnstile model, for p-pass streaming algorithms with $p > \log n$, this can be done with $O(\log n / \log \log n)$ additional passes w.h.p. using $O(\log n)$ ℓ_0-samplers [28] for each node $v \in V \setminus \{s\}$, and this costs $\tilde{O}(n)$ space. For $p \leq \log n$, the space bound is $\tilde{O}(n^2)$ and one can use $\tilde{O}(n)$ ℓ_0-samplers for each node, so this step can be done in one additional pass. Hence in the subsequent discussion we focus on computing the distance from s to each node $v \in V \setminus \{s\}$.

3.1 A Simple Deterministic Algorithm

We present a simple deterministic p-pass $\tilde{O}(n^2/p)$-space algorithm in the insertion-only model by an observation that every root-to-leaf path in a BFS tree cannot visit too many high-degree nodes (Lemma 8). Then, one can simulate the standard BFS algorithm efficiently layer-by-layer over high-degree nodes (Theorem 9).

▶ **Lemma 8.** *Let P be a root-to-leaf path in some BFS tree of an n-node connected simple undirected graph G. Then*

$$\sum_{x \in P} \deg_G(x) \leq 3n = O(n)$$

where $\deg_G(x)$ denotes the degree of x in G.

Proof. Suppose $P = x_1 x_2 \cdots x_k$ comprises k nodes. Observe that if x_i and x_j have $i \equiv j$ (mod 3), then x_i and x_j cannot share any neighbor node; otherwise P can be shorten, a contradiction. Thus, for each $c \in \{0, 1, 2\}$ the total contribution of all x_i's whose $i \equiv c$ (mod 3) to $\sum_{x_i \in P} \deg_G(x_i)$ is $O(n)$. Summing over all possible c gives the bound. ◀

We note that Lemma 8 is near-optimal. To see why, let $H = (V, E)$ where V is the union of disjoint sets $V_0, V_1, \ldots, V_k$ and $E = \{(x, y) : x \in V_i \text{ and } y \in V_j \text{ for any } i, j \text{ that } |i - j| \leq 1\}$. By setting $k = \lceil (n-1)/t \rceil$ for some parameter t, $|V_0| = 1$, $|V_i| = t$ for every $i \in [1, k-1]$, and $1 \leq |V_k| \leq t$, any BFS tree rooted at the node in V_0 has a root-to-leaf path Q of length k, and each node in $Q \cap (V_2 \cup V_3 \cup \ldots \cup V_{k-2})$ has degree $3t - 1$. Pick any t such that $k = \omega(1)$ and $t = \omega(1)$. We have $\sum_{x \in Q} \deg_H(x) = (3 - o(1))n$.

▶ **Theorem 9.** *Given an n-node connected simple undirected graph G with a distinguished node s, a BFS tree rooted at s can be found deterministically in p passes using $\tilde{O}(n^2/p)$ space for every $p \in [1, n]$ in the insertion-only model.*

Proof. Given a parameter k, our algorithm goes as follows. In the first pass, keep arbitrary n/k neighbors for each node $v \in G$ in memory and then use the in-memory edges to update the distance $\text{dist}(s, v)$ for each $v \in G$ by any single-source shortest path algorithm. The set of the in-memory edges is an invariant after the first pass. Hence, the memory usage is

$\tilde{O}(n^2/k)$. Then, in each of the subsequent passes, processing the edges (u, v) in the stream one by one, without keeping them in memory after the processing, if $\mathrm{dist}(s, u) + 1 < \mathrm{dist}(s, v)$ (resp. if $\mathrm{dist}(s, v) + 1 < \mathrm{dist}(s, u)$), then update $\mathrm{dist}(s, v)$ (resp. $\mathrm{dist}(s, u)$). After the edges in the stream are all processed, use the in-memory edges to update the distance $\mathrm{dist}(s, v)$ for each $v \in G$ again by any single-source shortest path algorithm but with initial distances. Our algorithm repeats until no distance has been updated in a single pass.

Observe a root-to-leaf path $P = sz_1 z_2 \cdots z_t$ in some BFS tree rooted at s. Suppose P contains exactly ℓ edges that appears only on tape, let them be $(z_{x_1}, z_{y_1}), \ldots, (z_{x_\ell}, z_{y_\ell})$ where $1 \le x_i < y_i \le x_{i+1} < y_{i+1} \le t$ for every $i \in [1, \ell - 1]$. Let $\mathrm{pred}_P(z_i)$ be the predecessor of z_i on P that is closest to z_i among nodes in $\{s\} \cup \{z_{y_j} : y_j < i\}$. By the definition of the above construction, it is assured that $\deg(z_{x_i}) \ge n/k$ for each $i \in [1, \ell]$. Thus by Lemma 8, $\ell = O(k)$. Then we appeal to the argument used for the analysis of Bellman-Ford algorithm [19, 6]. For every $i \in [1, t]$, if $i \notin \{y_1, y_2, \ldots, y_\ell\}$, $\mathrm{dist}(s, z_i)$ attains the minimum possible value at the same pass when $\mathrm{dist}(s, \mathrm{pred}_P(z_i))$ attains; otherwise $i = y_j$ for some $j \in [1, \ell]$, $\mathrm{dist}(s, y_j)$ attains the minimum possible value at most one pass after $\mathrm{dist}(s, x_j)$ attains. Hence, $O(k)$ passes suffices to compute $\mathrm{dist}(s, z_i)$ for all $i \in [1, t]$ and this argument applies to all root-to-leaf paths. Setting $k = p$ yields the desired bound. ◀

3.2 A More Efficient Randomized Algorithm

In this section, we prove Theorem 3. Our BFS algorithm is based on the following generic framework, which has been applied to finding shortest paths in the parallel and the distributed settings [11, 22, 27, 45]. Sample a set U of approximately k distinguished nodes such that each node $v \ne s$ joins U independently with probability k/n, and $s \in U$ with probability 1. By a Chernoff bound, $|U| = \tilde{\Theta}(k)$ with high probability. We will grow a local BFS tree of radius $\tilde{O}(n/k)$ from each node in U, and then we will construct the final BFS tree by combining them. We will rely on the following lemma, which first appeared in [45].

▶ **Lemma 10** ([45]). *Let s be a specified source node. Let U be a subset of nodes such that each node $v \ne s$ joins U with probability k/n, and s joins U with probability 1. For any given parameter $C \ge 1$, the following holds with probability $1 - n^{-\Omega(C)}$. For each node $t \ne s$, there is an s-t shortest path $P_{s,t}$ such that each of its $C(n \log n)/k$-node subpath P' satisfies $P' \cap U \ne \emptyset$.*

For notational simplicity, in subsequent discussion we write $h = C(n \log n)/k - 1 = \tilde{O}(n/k)$. Lemma 10 shows that for each node $t \in V \setminus \{s\}$,

$$\mathrm{dist}(s, t) = \min_{u \in U \cap N^h(t)} \mathrm{dist}(s, u) + \mathrm{dist}(u, t) \tag{2}$$

with probability $1 - n^{-\Omega(C)}$ where $N^h(v) = \{u : \mathrm{dist}(u, v) \le h\}$.

To see this, consider the s-t shortest path $P_{s,t}$ specified in Lemma 10. If the number of nodes in $P_{s,t}$ is less than h, then the above claim holds because $s \in U \cap N^h(t)$. Otherwise, Lemma 10 guarantees that there is a node $u \in P_{s,t} \cap U \cap N^h(t)$ with probability $1 - n^{-\Omega(C)}$. Using Equation (2), a BFS tree can be computed using the following steps.

1. Compute $\mathrm{dist}(u, v)$ for each $u \in U$ and $v \in U \cap N^h(u)$. Using this information, we can infer $\mathrm{dist}(s, u)$ for each $u \in U$.
2. Compute $\mathrm{dist}(s, t)$ for each $t \in V \setminus \{s\}$ by the formula $\mathrm{dist}(s, t) = \min_{u \in U \cap N^h(t)} \mathrm{dist}(s, u) + \mathrm{dist}(u, t)$.

In what follows, we show how to implement the above two steps in the streaming model, using $\tilde{O}(n + k^2)$ space and $\tilde{O}(n/k)$ passes. By a change of parameter $p = \tilde{O}(n/k)$, we obtain Theorem 3.

Step 1. To compute $\text{dist}(u,v)$ for each $u \in U$ and $v \in U \cap N^h(u)$, we let each $u \in U$ initiate a radius-h local BFS rooted at u. A straightforward implementation of this approach in the streaming model costs $h = \tilde{O}(n/k)$ passes and $O(n \cdot |U|) = \tilde{O}(nk)$ space, since we need to maintain $|U|$ search trees simultaneously.

We show that the space requirement can be improved to $\tilde{O}(n + k^2)$. Since we only need to learn the distances between nodes in U, we are allowed to forget distance information associated with nodes $v \notin U$ when it is no longer needed. Specifically, suppose we start the BFS computation rooted at $u \in U$ at the τ_uth pass, where τ_u is some number to be determined. For each $0 \le i \le h-1$, the induction hypothesis specifies that at the beginning of the $(\tau_u + i)$th pass, all nodes in $L_i(u) = \{v \in V : \text{dist}(u,v) = i\}$ have learned that $\text{dist}(u,v) = i$. During the $(\tau_u + i)$th pass, for each node $v \in V$ with $\text{dist}(u,v) > i$, we check if v has a neighbor in $L_i(u)$. If so, then we learn that $\text{dist}(u,v) = i+1$.

In the above BFS algorithm, if $\text{dist}(u,v) = i$ for some $0 \le i \le h-1$, then we learn the fact that $\text{dist}(u,v) = i$ during the $(\tau_u + i - 1)$th pass. Observe that such information is only needed during the next two passes. After the end of the $(\tau_u + i + 1)$th pass, for each $v \in V$ with $\text{dist}(u,v) = i$, we are allowed to forget that $\text{dist}(u,v) = i$. That is, v only needs to participate in the BFS computation rooted at u during these three passes $\{\tau_u + i - 1, \ \tau_u + i, \ \tau_u + i + 1\}$.

For each $u \in U$, we assign the starting time τ_u independently and uniformly at random from $\{1, 2, \ldots, h\}$. Lemma 11 shows that for each node $v \in V$ and for each pass $1 \le t \le 2h-1$, the number of BFS computations that involve v is $\tilde{O}(1)$. The idea of using random starting time to schedule multiple algorithms to minimize congestion can be traced back from [32]. Note that $\tau_u + \text{dist}(u,v) - 1 \le t \le \tau_u + \text{dist}(u,v) + 1$ is the criterion for v to participate in the BFS rooted at u during the tth pass.

▶ **Lemma 11.** *For each node v, and for each integer $1 \le t \le 2h-1$, with high probability, the number of nodes $u \in U$ such that $\tau_u + \text{dist}(u,v) - 1 \le t \le \tau_u + \text{dist}(u,v) + 1$ is at most $O(\max\{\log n, |U|/h\})$.*

Proof. Given two nodes $u \in U$ and $v \in V$, and a fixed number t, the probability that $\tau_u + \text{dist}(u,v) - 1 \le t \le \tau_u + \text{dist}(u,v) + 1$ is at most $3/h$. Let X be the total number of $u \in U$ such that $\tau_u + \text{dist}(u,v) - 1 \le t \le \tau_u + \text{dist}(u,v) + 1$. The expected value of X can be upper bounded by $\mu = |U| \cdot (3/h)$. By a Chernoff bound, with high probability, $X = O(\max\{\log n, |U|/h\})$. ◀

Recall that $|U| = \tilde{O}(k)$ with high probability, and $h = \tilde{O}(n/k)$. By Lemma 11, we only need $\lceil k^2/n \rceil \cdot \tilde{O}(1)$ space per each $v \in V$ to do the radius-h BFS computation from all nodes $u \in U$. That is, the space complexity is $\tilde{O}(n + k^2)$. To store the distance information $\text{dist}(u,v)$ for each $u \in U$ and $v \in U \cap N^h(u)$, we need $\tilde{O}(k^2)$ space. Thus, the algorithm for Step 1 costs $\tilde{O}(n + k^2)$ space. The number of passes is $2h - 1 = \tilde{O}(k)$.

In the insertion-only model, the implementation is straightforward. In the turnstile model, care has to be taken when implementing the above algorithm. We write $x = O(\max\{\log n, |U|/h\})$ to be the high probability upper bound on the number of BFS computation that a node participates in a single pass. We write $y = O(x \log n)$. Let $U_1, U_2, \ldots, U_y$ be random subsets of U such that each $u \in U$ joins each U_j with probability $1/x$, independently. Consider a node $v \in V$ and consider the rth pass. Let S be the subset of U such that $u \in S$ if $r = \tau_u + \text{dist}(u,v) - 1$, i.e., the BFS computation rooted at u hits v during the rth pass. We know that with high probability $|S| \le x$. By our choice of $U_1, U_2, \ldots, U_y$, we can infer that with high probability for each $u \in S$ there is at least one index j such that $S \cap U_j = \{u\}$.

To implement the rth pass in the turnstile model, each node $v \in V$ virtually maintains y edge set $Z_1, Z_2, \ldots, Z_y$. For each insertion (resp., deletion) of an edge $e = \{w, v\}$ satisfying $r = \tau_u + \text{dist}(u, w) - 2$ for some $u \in U_j$, we add (resp., remove) the edge from the set Z_j. After processing the entire data stream, we take one edge out of each edge set $Z_1, Z_2, \ldots, Z_y$. In view of the above discussion, it suffices to only consider these edges when we grow the BFS trees. This can be implemented using y ℓ_0-samplers per each node $v \in V$, and the space complexity is still $\tilde{O}(ny) = \tilde{O}(n + k^2)$.

Step 2.   At this moment we have computed $\text{dist}(s, u)$ for each $u \in U$. Now we need to compute $\text{dist}(s, t)$ for each $t \in V \setminus \{s\}$ by the formula $\text{dist}(s, t) = \min_{u \in U \cap N^h(t)} \text{dist}(s, u) + \text{dist}(u, t)$.

In the insertion-only model, this task can be solved using h iterations of Bellman-Ford steps. Initially, $d_0(v) = \text{dist}(s, v)$ for each $v \in U$, and $d_0(v) = \infty$ for each $v \in V \setminus U$. During the ith pass, we do the update $d_i(v) \leftarrow \min\{d_{i-1}(v), 1 + \min_{u \in N(v)} d_{i-1}(u)\}$. By Equation (2), we can infer that $d_h(t) = \text{dist}(s, t)$ for each $t \in V$. A straightforward implementation of this procedure costs $\tilde{O}(n)$ space and $h = \tilde{O}(n/k)$ passes.

In the turnstile model, we can solve this task by growing a radius-h BFS tree rooted at u, for each $u \in U$, as in Step 1. During the process, each node $v \in V$ maintains a variable $d(v)$ which serves as the estimate of $\text{dist}(s, v)$. Initially, $d(v) \leftarrow \infty$. When the partial BFS tree rooted at $u \in U$ hits v, we update $d(v)$ to be the minimum of the current value of $d(v)$ and $\text{dist}(s, u) + \text{dist}(u, v)$. At the end of the process, we have $d(v) = \min_{u \in U \cap N^h(t)} \text{dist}(s, u) + \text{dist}(u, v) = \text{dist}(s, v)$ for each $v \in V$. This costs $\tilde{O}(n + k^2)$ space and $\tilde{O}(n/k)$ passes in view of the analysis of Step 1.

3.3   Extensions

In this section we consider the problem of solving c instances of BFS simultaneously for some $c \leq n$ and a simpler problem of computing the pairwise distance between the c given nodes. Both of these problems can be solved via a black box application of Theorem 3. In this section we show that it is possible to obtain better upper bounds.

▶ **Theorem 12.** *Given an n-node undirected graph G, for any given parameters $1 \leq c \leq k \leq n$, the pairwise distances between all pairs of nodes in a given set of c nodes in G can be computed with probability $1 - 1/n^{\Omega(1)}$ using $\tilde{O}(n/k)$ passes and $\tilde{O}(n + k^2)$ space in the turnstile model.*

Proof. Let S be the input node set of size c. Consider the modified Step 1 of our algorithm where each $s \in S$ is included in U with probability 1. Since $|S| = c \leq k$, we still have $|U| = \tilde{O}(k)$ with high probability. Recall that Step 1 of our algorithm calculates $\text{dist}(u, v)$ for each $u \in U$ and $v \in U \cap N^h(u)$ in $\tilde{O}(n + k^2)$ space and $\tilde{O}(n/k)$ passes. Applying Equation (2) for each $s \in U$, we obtain the pairwise distances between all pairs of nodes in U, which includes S as a subset. There is no need to do Step 2.   ◀

For example, if $c = n^{1/2}$, then Theorem 12 implies that we can compute the pairwise distances between all pairs of nodes in a given set of c nodes in $\tilde{O}(n)$ space and $\tilde{O}(n^{1/2})$ passes.

▶ **Theorem 13.** *Given an n-node undirected graph G, for any given parameters $1 \leq c \leq k \leq n$, one can solve c instances of BFS with probability $1 - 1/n^{\Omega(1)}$ using $\tilde{O}(n/k)$ passes and $\tilde{O}(cn + k^2)$ space in the turnstile model.*

Proof. Let S be the node set of size c corresponding to the roots of the c BFS instances. Consider the following modifications to our BFS algorithm.

Same as the proof of Theorem 12, in Step 1, include each $s \in S$ in U with probability 1. The modified Step 1 still takes $\tilde{O}(n + k^2)$ space and $\tilde{O}(n/k)$ passes, and it outputs the pairwise distances between all pairs of nodes in U.

Now consider Step 2. In the insertion-only model, remember that a BFS tree rooted at a node $s \in S$ can be constructed in $O(n)$ space and $h = \tilde{O}(n/k)$ passes using h iterations of Bellman-Ford steps. The cost of constructing all c BFS trees is then $O(cn)$ space and $\tilde{O}(n/k)$ passes.

In the turnstile model, we can also use the strategy of growing a radius-h BFS tree rooted at u, for each $u \in U$. During the process, each node $v \in V$ maintains c variables serving as the estimates of $\text{dist}(s, v)$, for all $s \in S$. The complexity of growing radius-h BFS trees is still $\tilde{O}(n + k^2)$ space and $\tilde{O}(n/k)$ passes. The extra space cost for maintaining these cn variables is $O(cn)$. ◀

For example, if $c = n^{1/3}$, then Theorem 13 implies that we can solve c instances of BFS in $\tilde{O}(n^{4/3})$ space and $\tilde{O}(n^{1/3})$ passes. Note that the space complexity of $\tilde{O}(n^{4/3})$ is necessary to output $c = n^{1/3}$ BFS trees.

Theorem 13 immediately gives the following corollary.

▶ **Corollary 14.** *Given an n-node connected undirected graph G with unweighted edges and a c-node subset S of G, for any given parameters $1 \le c \le k \le n$, finding a Steiner tree in G that spans S can be approximated to within a factor of 2 with probability $1 - 1/n^{\Omega(1)}$ using $\tilde{O}(n/k)$ passes and $\tilde{O}(cn + k^2)$ space in the turnstile model.*

Note that if we do not need to construct a Steiner tree, and only need to approximate the size of an optimal Steiner tree, then Theorem 12 can be used in place of Theorem 13.

3.4 Diameter Approximation

It is well-known that the maximum distance label in a BFS tree gives a 2-approximation of diameter. We show that it is possible to improve the approximation ratio to nearly 1.5 without sacrificing the space and pass complexities.

Roditty and Williams [40] showed that a nearly 1.5-approximation of diameter can be computed with high probability as follows.

1. Let S_1 be a node set chosen by including each node $v \in V$ to S_1 with probability $p = (\log n)/\sqrt{n}$ independently. Perform a BFS from each node $v \in S_1$.
2. Let $v^\star$ be a node chosen to maximize $\text{dist}(v^\star, S_1)$. Break the tie arbitrarily. Perform a BFS from $v^\star$.
3. Let S_2 be the node set consisting of the $\sqrt{n}$ nodes closest to $v^\star$, where ties are broken arbitrarily. Perform a BFS from each node $v \in S_2$.

Let D^* be the maximum distance label ever computed during the BFS computations in the above procedure. Roditty and Williams [40] proved that D^* satisfies that $\lfloor 2D/3 \rfloor \le D^* \le D$, where D is the diameter of G.

The algorithm of Roditty and Williams [40] can be implemented in the streaming model by applying Theorem 13 with $c = \tilde{O}(\sqrt{n})$, but we can do better. Note that when we perform BFS from the nodes in S_1 and S_2, it is not necessary to store the entire BFS trees. For example, in order to select v^*, we only need to let each node v know $\text{dist}(v, S_1)$, which is the maximum distance label of v in all BFS trees computed in Step 1. Therefore, the $O(cn)$ term in the space complexity of Theorem 13 can be improved to $O(n)$. That is, the space and pass complexities are the same as the cost for computing a *single* BFS tree using Theorem 3. We conclude the following theorem.

▶ **Theorem 15.** *Given an n-node connected undirected graph G, a diameter approximation D^* satisfying $\lfloor 2D/3 \rfloor \leq D^* \leq D$, where D is the diameter of G, can be computed with probability $1 - 1/n^{\Omega(1)}$ in p passes using $\tilde{O}((n/p)^2)$ space, for each $1 \leq p \leq \tilde{O}(\sqrt{n})$ in the turnstile model.*

4   Depth-First Search

A straightforward implementation of the naive DFS algorithm in the streaming model costs either $n - 1$ passes with $\tilde{O}(n)$ space or a single pass with $O(n^2)$ space. Khan and Mehta [30] recently showed that it is possible to obtain a smooth tradeoff between the two extremes. Specifically, they designed an algorithm that requires at most $\lceil n/k \rceil$ passes using $\tilde{O}(nk)$ space, where k is any positive integer. Furthermore, for the case the height h of the computed DFS tree is small, they further decrease the number of passes to $\lceil h/k \rceil$. In Section 4.1, we will provide a very simple alternative proof of this result, via sparse certificates for k-node-connectivity.

In the worst case, the "space $\times$ number of passes" of the algorithms of Khan and Mehta [30] is still $\tilde{O}(n^2)$. In Section 4.2, we will show that it is possible to improve this upper bound asymptotically when the number of passes p is super-constant. Specifically, for any parameters $1 \leq s \leq k \leq n$, we obtain the following DFS algorithms.

- A deterministic algorithm using $\tilde{O}((n/k) + (k/s))$ passes and $\tilde{O}(ns)$ space in the insertion-only model. After balancing the parameters, the space complexity is $\tilde{O}(n^2/p^2)$ for p-pass algorithms, for each $1 \leq p \leq \tilde{O}(\sqrt{n})$.
- A randomized algorithm using $\tilde{O}((n/k) + (k/s))$ passes and $\tilde{O}(ns^2)$ space in the turnstile model. After balancing the parameters, the space complexity is $\tilde{O}(n^3/p^4)$ for p-pass algorithms, for each $1 \leq p \leq \tilde{O}(\sqrt{n})$.

4.1   A Simple DFS Algorithm

In this section, we present a simple alternative proof of the result of Khan and Mehta [30] that a DFS tree can be constructed in $\lceil h/k \rceil$ passes using $\tilde{O}(nk)$ space, for any given parameter k, where h is the height of the computed DFS tree.

Sparse Certificate for s-Node-Connectivity.   A *strong s-VC certificate* of a graph H is its subgraph K such that for any supergraph G of H, for every pair of nodes $s^*, t^* \in G$, if they are c-node-connected in G, then they are c'-node-connected for some $c' \geq \min\{s, c\}$ in the graph obtained from G by replacing its subgraph H with K. A *sparse* strong s-VC certificate of the graph G is exactly what we need here. Eppstein, Galil, Italiano, and Nissenzweig [12] showed that such a sparse subgraph of $O(ns)$ edges can be computed in a single pass with $\tilde{O}(ns)$ space *deterministically* in the insertion-only model. In the turnstile model, Guha, McGregor, and Tench [25] showed that such a sparse subgraph of $\tilde{O}(ns^2)$ edges can be computed with high probability in a single pass using $\tilde{O}(ns^2)$ space. This result can be inferred from Theorem 8 of [25] with $\epsilon = \Theta(1/s)$. In [25] the analysis only considers the case $G = H$, but it is straightforward to extend the analysis to incorporate any supergraph G of H.

If the subgraph K of the graph H satisfies the above requirement for the special case of $G = H$, then K is said to be a *s-VC certificate* of H. Our simple DFS algorithm relies on this tool.

▶ **Lemma 16.** *Suppose K is a $(k+1)$-VC certificate of H. Let T be any DFS tree of K. Consider any two nodes u and v such that the least common ancestor w of u and v are within the top k layers of T. If $w \neq u$ and $w \neq v$, then u and v are not adjacent in H.*

Proof. Suppose u, v, and w violate the statement of the lemma. That is, u and v are adjacent in H. Since T is a DFS tree, u and v are not adjacent in K, and each path connecting u and v must pass through a node that is a common ancestor of u and v. Let c_H (resp., c_K) be the u-v node-connectivity in H (resp., K). The above discussion implies that $c_K \leq k$ and $c_H \geq c_K + 1$, contradicting the assumption that K is a $(k+1)$-VC certificate of H.   ◄

Algorithm.   Using Lemma 16, we can construct a DFS tree of G recursively as follows. Pick K as a $(k+1)$-VC certificate of G. Compute a DFS tree T of K. Let T' be the tree induced by the top $k+1$ layers of of T. Let $v_1, v_2, \ldots, v_z$ be the leaves of T'. Denote S_i as the set of descendants of v_i in T, including v_i. By Lemma 16, there exists no edge in G that crosses two distinct sets S_i and S_j. For each $1 \leq i \leq z$, we recursively find a DFS tree T_i of the subgraph of G induced by S_i rooted at v_i. By the above observation, we can obtain a valid DFS tree of G by appending $T_1, T_2, \ldots, T_z$ to T'.

Analysis.   If the height of the final DFS tree is h, then the depth of the recursion is at most $\lceil h/k \rceil$. The cost for computing a $(k+1)$-VC certificate is 1 pass and $\tilde{O}(nk)$ space, and the resulting subgraph K has $O(nk)$ edges. Therefore, the total number of passes is at most $\lceil h/k \rceil$, and the overall space complexity is $\tilde{O}(nk)$.

4.2   Streaming Implementation of the Algorithm of Aggarwal and Anderson

The bounds of Theorem 4 are attained via an implementation of the parallel DFS algorithm of Aggarwal and Anderson [2] in the streaming model, with the help of various tools, including the strong sparse certificates for s-node-connectivity described above. Due to the page limit, we only provide a sketch of the proof. The complete proof will be presented in the full version of this paper.

At a high level, the DFS algorithm of Aggarwal and Anderson [2] works as follows. Start with a maximal matching, and then merge these length-1 paths iteratively into a constant number of node-disjoint paths such that the number of nodes not in any path is at most $|V|/2$. The algorithm then constructs the initial segment of the DFS tree from these paths. Each remaining connected component is solved recursively. The final DFS tree is formed by appending the DFS trees of recursive calls to the initial segment.

The bottleneck of this DFS algorithm is a task called MaximalPaths which is a variant of the maximal node-disjoint paths problem between a set of source nodes S and a set of sink nodes T. In this variant, each member of S is a path instead of a node. Goldberg, Plotkin, and Vaidya [24] gave a parallel algorithm for this problem. Their algorithm has two phases. For any given parameter k, they showed that after k iterations of the algorithm of the first phase, the number of sources in S that are still *active* is at most n/k. These remaining active sources are processed one-by-one in the second phase. Using this approach with $k = \sqrt{n}$, MaximalPaths can be solved in the streaming model with $\tilde{O}(\sqrt{n})$ passes and $\tilde{O}(n)$ space. To further reduce the pass complexity, we apply the sparse certificates for s-node-connectivity of Eppstein, Galil, Italiano, and Nissenzweig [12] and Guha, McGregor, and Tench [25], which allow us to process the remaining active sources in batches. In the insertion-only model, we obtain a deterministic p-pass algorithm with space complexity $\tilde{O}(n^2/p^2)$, for each $1 \leq p \leq \tilde{O}(\sqrt{n})$. For the more challenging turnstile model, we obtain a randomized algorithm with a somewhat worse space complexity of $\tilde{O}(n^3/p^4)$.

5 Single-Pass Lower Bounds

In this section, we use the lower bound of the 1-way randomized communication complexity for the INDEX problem [1] to show the single-pass space lower bound for computing approximate MLST to within an additive error k. This gives a complementary result for Theorem 1.

▶ **Theorem 17.** *In the insertion-only model, given a connected n-node simple undirected graph G, computing a spanning tree of G that has at least $\mathrm{leaf}(G) - k$ leaves for any $k \in [1, (n-5)/4]$ requires $\Omega(n^2/k^2)$ bits for any single-pass randomized streaming algorithm that can succeed with probability at least $2/3$.*

Proof. We begin with a reduction from an n^2-bit instance of the INDEX problem to computing a spanning tree of $(2n+3)$-node graph G with $\mathrm{leaf}(G)$ leaves for any $n \geq 1$. Given Alice's input in the INDEX problem, i.e. a bit-array of length n^2, we construct an n by n bipartite graph H, as part of G, in which edge (x_i, y_j) for every $i, j \in [1, n]$ corresponds to the $((i-1)n + j)$-th bit in Alice's array. Then, given Bob's input, a tuple (i, j) for some $i, j \in [1, n]$, we construct the remaining part of G by adding three additional nodes s, t, and ℓ, and

- connecting an edge from s to z for every node $z \neq y_j$ in H, and
- adding edge (ℓ, x_i), (s, t), and (t, y_j).

It clear that G is connected and has

$$\mathrm{leaf}(G) = \begin{cases} 2n + 1 & \text{if } (x_i, y_j) \in H \\ 2n & \text{otherwise} \end{cases}$$

Thus, having a single-pass streaming algorithm to compute $\mathrm{leaf}(G)$ suffices to decide the n^2-bit instance of the INDEX problem, i.e. for Bob to tell what the $((i-1)n + j)$-th bit in Alice's array is. This requires $\Omega(n^2)$ bits. To obtain the hardness result for MLST with additive error k for any $k \geq 1$, one can duplicate $H \cup \{\ell, t\}$ into $(k+1)$ copies and let the copies share the same s, so G is connected, has $(k+1)(2n+2) + 1$ nodes, and has

$$\mathrm{leaf}(G) = \begin{cases} (2n+1)(k+1) & \text{if } (x_i, y_j) \in H \\ 2n(k+1) & \text{otherwise} \end{cases}$$

Hence, having a single-pass streaming algorithm to compute $\mathrm{leaf}(G)$ for G of $(k+1)(2n+2) + 1$ nodes to within an additive error k suffices to decide the n^2-bit INDEX problem. Replace $(k+1)(2n+2) + 1 = n'$ and $n^2 = \Omega((n'/k)^2)$ yields the desired bound. ◀

6 Conclusion

In this paper, we devised semi-streaming algorithms for spanning tree computations, including max-leaf spanning trees, BFS trees, and DFS trees. For max-leaf spanning trees, despite that any streaming algorithm requires $\Omega(n^2)$ space to compute the exact solution, we show how to compute a $(1 + \varepsilon)$-approximation using a single pass and $\tilde{O}(n)$ space, albeit in super-polynomial time. For BFS trees and DFS trees, we show how to compute them using $O(\sqrt{n})$ passes and $\tilde{O}(n)$ space, and offer a smooth tradeoff between pass complexity and space usage.

The pass complexities of our algorithms for BFS trees and DFS trees are still far from the known lower bounds, $\omega(1)$ passes for BFS trees [17] and the trivial 1 pass for DFS trees. It is unclear whether our upper bounds can be further reduced or the known lower bounds can be improved. We leave closing the gap to future work.

──────── **References** ────────

1   Farid M. Ablayev. Lower bounds for one-way probabilistic communication complexity and their application to space complexity. *Theor. Comput. Sci.*, 157(2):139–159, 1996.

2   Alok Aggarwal and Richard J. Anderson. A random NC algorithm for depth first search. *Combinatorica*, 8(1):1–12, March 1988.

3   Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Analyzing graph structure via linear measurements. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2012, Kyoto, Japan, January 17-19, 2012*, pages 459–467, 2012.

4   Sepehr Assadi, Yu Chen, and Sanjeev Khanna. Sublinear algorithms for $(\Delta + 1)$ vertex coloring. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019*, pages 767–786, 2019.

5   Jesse Beisegel. Characterising AT-free graphs with BFS. In *Graph-Theoretic Concepts in Computer Science - 44th International Workshop, WG 2018, Cottbus, Germany, June 27-29, 2018, Proceedings*, pages 15–26, 2018.

6   Richard Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16(1):87–90, 1958.

7   Vladimir Braverman, Rafail Ostrovsky, and Dan Vilenchik. How hard is counting triangles in the streaming model? In *Automata, Languages, and Programming - 40th International Colloquium, ICALP 2013, Riga, Latvia, July 8-12, 2013, Proceedings, Part I*, pages 244–254, 2013.

8   Joseph Cheriyan and Ramakrishna Thurimella. Algorithms for parallel k-vertex connectivity and sparse certificates. In *Proceedings of the Twenty-third Annual ACM Symposium on Theory of Computing*, STOC '91, pages 391–401. ACM, 1991.

9   Miroslav Chlebík and Janka Chlebíková. Approximation hardness of dominating set problems in bounded degree graphs. *Inf. Comput.*, 206(11):1264–1275, 2008.

10   Derek G. Corneil, Feodor F. Dragan, and Ekkehard Köhler. On the power of BFS to determine a graph's diameter. *Networks*, 42(4):209–222, 2003.

11   M. Elkin. Distributed exact shortest paths in sublinear time. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 757–770, New York, NY, USA, 2017. ACM.

12   David Eppstein, Zvi Galil, Giuseppe F. Italiano, and Amnon Nissenzweig. Sparsification – a technique for speeding up dynamic graph algorithms. *JACM*, 44(5):669–696, 1997.

13   Shimon Even and Robert Endre Tarjan. Computing an st-numbering. *Theoretical Computer Science*, 2(3):339–344, 1976.

14   Martin Farach-Colton, Tsan-sheng Hsu, Meng Li, and Meng-Tsung Tsai. Finding articulation points of large graphs in linear time. In *Algorithms and Data Structures - 14th International Symposium, WADS 2015, Victoria, BC, Canada, August 5-7, 2015. Proceedings*, pages 363–372, 2015.

15   Martin Farach-Colton and Meng-Tsung Tsai. Tight approximations of degeneracy in large graphs. In *LATIN 2016: Theoretical Informatics - 12th Latin American Symposium, Ensenada, Mexico, April 11-15, 2016, Proceedings*, pages 429–440, 2016.

16   Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. On graph problems in a semi-streaming model. *Theor. Comput. Sci.*, 348(2-3):207–216, 2005.

17   Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. Graph distances in the data-stream model. *SIAM J. Comput.*, 38(5):1709–1727, 2008.

18   T. Fleiner and G. Wiener. Coloring signed graphs using DFS. *Optimization Letters*, 10(4):865–869, April 2016.

19   L.R. Ford. *Network Flow Theory*. Paper P. Rand Corporation, 1956.

20   Giulia Galbiati, Francesco Maffioli, and Angelo Morzenti. A short note on the approximability of the maximum leaves spanning tree problem. *Inf. Process. Lett.*, 52(1):45–49, 1994.

21   Rajiv Gandhi, Mohammad Taghi Hajiaghayi, Guy Kortsarz, Manish Purohit, and Kanthi K. Sarpatwar. On maximum leaf trees and connections to connected maximum cut problems. *Inf. Process. Lett.*, 129:31–34, 2018.

22   Mohsen. Ghaffari and Jason. Li. Improved distributed algorithms for exact shortest paths. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2018.

23   Mohsen Ghaffari and Merav Parter. Near-optimal distributed DFS in planar graphs. In *31st International Symposium on Distributed Computing, DISC 2017, October 16-20, 2017, Vienna, Austria*, pages 21:1–21:16, 2017.

24   A. V. Goldberg, S. A. Plotkin, and P. M. Vaidya. Sublinear-time parallel algorithms for matching and related problems. *JALG*, 14(2):180–213, 1993.

25   Sudipto Guha, Andrew McGregor, and David Tench. Vertex and hyperedge connectivity in dynamic graph streams. In *Proceedings of the 34th ACM Symposium on Principles of Database Systems, PODS 2015, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 241–247, 2015.

26   Mohammad Taghi Hajiaghayi, Guy Kortsarz, Robert MacDavid, Manish Purohit, and Kanthi K. Sarpatwar. Approximation algorithms for connected maximum cut and related problems. In *Algorithms - ESA 2015 - 23rd Annual European Symposium, Patras, Greece, September 14-16, 2015, Proceedings*, pages 693–704, 2015.

27   Chien-Chung Huang, Danupon Nanongkai, and Thatchaphol Saranurak. Distributed exact weighted all-pairs shortest paths in $\tilde{O}(n^{5/4})$ rounds. In *IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 168–179, 2017.

28   Hossein Jowhari, Mert Sağlam, and Gábor Tardos. Tight bounds for ℓ_p samplers, finding duplicates in streams, and related problems. In *Proceedings of the Thirtieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '11, pages 49–58, New York, NY, USA, 2011. ACM.

29   Michael Kapralov, Yin Tat Lee, Cameron Musco, Christopher Musco, and Aaron Sidford. Single pass spectral sparsification in dynamic streams. *SIAM J. Comput.*, 46(1):456–477, 2017.

30   Shahbaz Khan and Shashank Mehta. Depth first search in the semi-streaming model. In *36th International Symposium on Theoretical Aspects of Computer Science (STACS 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.

31   Daniel J. Kleitman and Douglas B. West. Spanning trees with many leaves. *SIAM J. Discrete Math.*, 4(1):99–106, 1991.

32   F. T. Leighton, B. M. Maggs, and S. B. Rao. Packet routing and job-shop scheduling in $O(\text{Congestion} + \text{Dilation})$ steps. *Combinatorica*, 14(2):167–186, 1994.

33   R. Lipton and R. Tarjan. A separator theorem for planar graphs. *SIAM Journal on Applied Mathematics*, 36(2):177–189, 1979.

34   Hsueh-I Lu and R. Ravi. The power of local optimization: Approximation algorithms for maximum-leaf spanning tree. In *In Proceedings, Thirtieth Annual Allerton Conference on Communication, Control and Computing*, pages 533–542, 1992.

35   Hsueh-I Lu and R. Ravi. Approximating maximum leaf spanning trees in almost linear time. *J. Algorithms*, 29(1):132–141, 1998.

36   Andrew McGregor, David Tench, Sofya Vorotnikova, and Hoa T. Vu. Densest subgraph in dynamic graph streams. In *Mathematical Foundations of Computer Science 2015 - 40th International Symposium, MFCS 2015, Milan, Italy, August 24-28, 2015, Proceedings, Part II*, pages 472–482, 2015.

37   S. Muthukrishnan. Data streams: Algorithms and applications. *Found. Trends Theor. Comput. Sci.*, 1(2):117–236, August 2005.

38   Hiroshi Nagamochi and Toshihide Ibaraki. A linear-time algorithm for finding a sparse k-connected spanning subgraph of a k-connected graph. *Algorithmica*, 7(5&6):583–596, 1992.

39   Jelani Nelson and Huacheng Yu. Optimal lower bounds for distributed and streaming spanning forest computation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete*

Algorithms, SODA 2019, San Diego, California, USA, January 6-9, 2019, pages 1844–1860, 2019.

40 Liam Roditty and Virginia Vassilevska Williams. Fast approximation algorithms for the diameter and radius of sparse graphs. In *Proceedings 45th ACM Symposium on Theory of Computing (STOC)*, pages 515–524, 2013.

41 Jens M. Schmidt. A simple test on 2-vertex- and 2-edge-connectivity. *Inf. Process. Lett.*, 113(7):241–244, 2013.

42 Roberto Solis-Oba, Paul S. Bonsma, and Stefanie Lowski. A 2-approximation algorithm for finding a spanning tree with maximum number of leaves. *Algorithmica*, 77(2):374–388, 2017.

43 Xiaoming Sun and David P. Woodruff. Tight bounds for graph problems in insertion streams. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2015, August 24-26, 2015, Princeton, NJ, USA*, pages 435–448, 2015.

44 Robert Endre Tarjan. A note on finding the bridges of a graph. *Inf. Process. Lett.*, 2(6):160–161, 1974.

45 J. D. Ullman and M. Yannakakis. High-probability parallel transitive-closure algorithms. *SIAM Journal on Computing*, 20(1):100–125, 1991.