Streaming Algorithms for Geometric Steiner Forest

Artur Czumaj[*]
University of Warwick

Shaofeng H.-C. Jiang[†]
Peking University

Robert Krauthgamer[‡]
Weizmann Institute of Science

Pavel Veselý[§]
Charles University

November 4, 2021

Abstract

We consider an important generalization of the Steiner tree problem, the *Steiner forest problem*, in the Euclidean plane: the input is a multiset $X \subseteq \mathbb{R}^2$, partitioned into k color classes $C_1, C_2, \ldots, C_k \subseteq X$. The goal is to find a minimum-cost Euclidean graph G such that every color class C_i is connected in G. We study this Steiner forest problem in the streaming setting, where the stream consists of insertions and deletions of points to X. Each input point $x \in X$ arrives with its color $\mathsf{color}(x) \in [k]$, and as usual for dynamic geometric streams, the input points are restricted to the discrete grid $\{0, \ldots, \Delta\}^2$.

We design a single-pass streaming algorithm that uses $\mathrm{poly}(k \cdot \log \Delta)$ space and time, and estimates the cost of an optimal Steiner forest solution within ratio arbitrarily close to the famous Euclidean Steiner ratio α_2 (currently $1.1547 \leq \alpha_2 \leq 1.214$). This approximation guarantee matches the state of the art bound for streaming Steiner tree, i.e., when $k = 1$. Our approach relies on a novel combination of streaming techniques, like sampling and linear sketching, with the classical Arora-style dynamic-programming framework for geometric optimization problems, which usually requires large memory and has so far not been applied in the streaming setting.

We complement our streaming algorithm for the Steiner forest problem with simple arguments showing that any finite approximation requires $\Omega(k)$ bits of space.

[*] Research partially supported by the Centre for Discrete Mathematics and its Applications (DIMAP), by a Weizmann-UK Making Connections Grant, by an IBM Faculty Award, and by EPSRC award EP/V01305X/1. Email: `A.Czumaj@warwick.ac.uk`

[†] Part of this work was done when the author was at the Weizmann Institute of Science and Aalto University. Email: `shaofeng.jiang@pku.edu.cn`

[‡] Work partially supported by ONR Award N00014-18-1-2364, by the Israel Science Foundation grant #1086/18, by a Weizmann-UK Making Connections Grant, and by a Minerva Foundation grant. Email: `robert.krauthgamer@weizmann.ac.il`

[§] Part of this work was done when the author was at the University of Warwick. Partially supported by European Research Council grant ERC-2014-CoG 647557, by a Weizmann-UK Making Connections Grant, by GA ČR project 19-27871X, and by Charles University project UNCE/SCI/004. Email: `vesely@iuuk.mff.cuni.cz`

1 Introduction

We study combinatorial optimization problems in dynamic geometric streams, in the classical framework introduced by [Ind04]. In this setting, focusing on low dimension $d = 2$, the input point set is presented as a stream of insertions and deletions of points restricted to the discrete grid $[\Delta]^2 := \{0, \dots, \Delta\}^2$. Geometric data is very common in applications, and has been a central object of algorithmic study, from different computational paradigms (like data streams, property testing and distributed/parallel computing) to different application domains (like sensor networks and scientific computing). Research on geometric streaming algorithms has been very fruitful, and in particular, streaming algorithms achieving $(1+\varepsilon)$-factor estimation (i.e., approximation of the optimal value) have been obtained for fundamental geometric problems, such as k-clustering [FS05, BFL+17, HSYZ19], facility location [LS08, CLMS13], and minimum spanning tree (MST) [FIS08].

Despite this significant progress, some similarly looking problems are still largely open. Specifically, for the TSP and Steiner tree problems, which are the cornerstone of combinatorial optimization, it is a major outstanding question (see, e.g., [Soh12]) whether a streaming algorithm can match the $(1+\varepsilon)$-approximation known for the offline setting [Aro98, Mit99]. In fact, the currently best streaming algorithms known for TSP and Steiner tree only achieve $O(1)$-approximation, and follow by a trivial application of the MST streaming algorithm.

While MST is closely related to TSP and Steiner tree – their optimal values are within a constant factor of each other – it seems unlikely that techniques built around MST could achieve $(1+\varepsilon)$-approximation for either problem. Indeed, even in the offline setting, the only approach known to achieve $(1+\varepsilon)$-approximation for TSP and/or Steiner tree relies on a framework devised independently by Arora [Aro98] and Mitchell [Mit99], that combines geometric decomposition (e.g., a randomly shifted quad-tree) and dynamic programming. These two techniques have been used *separately* in the streaming setting in the past: quad-tree decomposition in [ABIW09, AIK08, Cha02, CLMS13, FIS08, FS05, IT03, LS08] and dynamic programming, mainly for string processing problems, in [BZ16, CGK16, CFH+21, EJ15, GJKK07, SS13, SW07]. However, we are not aware of any successful application of the Arora/Mitchell framework, which combines these two approaches, for any geometric optimization problem whatsoever.

We make an important step towards better understanding of these challenges by developing new techniques that *successfully adapt the Arora/Mitchell framework to streaming.* To this end, we consider a generalization of Steiner tree, the classical **Steiner Forest Problem (SFP)**. In this problem (also called *Generalized Steiner tree*, see, e.g., [Aro98]), the input is a multiset of n *terminal* points $X \subseteq [\Delta]^2$, partitioned into k *color classes* $X = C_1 \sqcup \dots \sqcup C_k$, presented as a dynamic stream. In addition, apart from the coordinates of the point $x \in X$, its color $\mathsf{color}(x) \in [k]$ is also revealed upon its arrival in the stream[1]. The goal is to find a minimum-cost Euclidean graph G such that every color class C_i is connected in G. Observe that the Steiner tree problem is a special case of SFP in which all terminal points should be connected (i.e., $k = 1$). Similarly to the Steiner tree problem, a solution to SFP may use points other than X; those points are called *Steiner points.*

Remark. In the literature, the term SFP sometimes refers to the *special case* where *each color class contains only a pair of points*, i.e., each $C_i = \{s_i, t_i\}$, see, e.g., [BH12, BHM11, BKM15, CHJ18, GK15]. It is not difficult to see (see [Sch16]) that one can reduce one problem into another in the standard setting of *offline algorithms.* The special case of pairs is often simpler to present and does not restrict algorithmic generality for offline algorithms, even though the setting considered here is more natural for applications (cf. [MW95]). Nevertheless, an important difference that we explore

[1] The points are arriving and leaving in an arbitrary order; there is no requirement that each color arrives in a batch, i.e., that its points are inserted/deleted consecutively in the stream; we discuss this special case in Section 6.

is that the definition used here allows for better parameterization over the number of colors k.

Background. While one might hope for a streaming algorithm for SFP with $o(k)$ space, we observe that *this task is impossible*, even in the one-dimensional case. In Theorem 4.1, we present a reduction that creates instances of SFP in $\mathbb{R}$ such that every streaming algorithm achieving any finite approximation ratio for SFP must use $\Omega(k)$ bits of space. This holds even for insertion-only algorithms, and even if all color classes are of size at most 2, and thus the input size is $n \leq 2k$.

Even for $k = 1$, which is the famous Steiner tree problem, the only known streaming algorithm is to estimate the cost of a minimum spanning tree (MST) and report it as an estimate for SFP. It is useful to recall here the *Steiner ratio* α_d, defined as the supremum over all point sets $X \subseteq \mathbb{R}^d$, of the ratio between the cost of an MST and that of an optimal Steiner tree. The famous Steiner ratio Gilbert-Pollak Conjecture [GP68] speculates that $\alpha_2 = \frac{2}{\sqrt{3}} \approx 1.1547$, but the best upper bound to date is only that $\alpha_2 \leq 1.214$ [CG85]. It follows that employing the streaming algorithm of Frahling, Indyk, and Sohler [FIS08], which $(1+\varepsilon)$-approximates the MST cost using space $\mathrm{poly}(\varepsilon^{-1}\log\Delta)$, immediately yields a streaming algorithm that $(\alpha_2+\varepsilon)$-approximates the Steiner tree cost, with the same space bound.

1.1 Our Contribution

Our main result, stated in Theorem 1.1, is a *space and time* efficient, *single-pass* streaming algorithm that estimates the optimal *cost* OPT for SFP within $(\alpha_2+\varepsilon)$ factor. Our space bound is nearly optimal in terms of the dependence in k, since any finite approximation for SFP requires space $\Omega(k)$ (Theorem 4.1), and our ratio matches the state of the art even for the special case $k = 1$.

Theorem 1.1 (Informal version of Theorem 3.1)**.** *For any integers $k, \Delta \geq 1$ and any fixed $\varepsilon > 0$, one can with high probability $(\alpha_2+\varepsilon)$-approximate the SFP cost of an input $X \subseteq [\Delta]^2$ presented as a dynamic geometric stream, using space and query and update times all bounded by* $\mathrm{poly}(k \cdot \log\Delta)$.

We notice that while the algorithm in Theorem 1.1 returns only an approximate cost of the optimal solution and it cannot return the entire approximate solution (since the output is of size $\Omega(n)$), an additional desirable feature of our algorithm in Theorem 1.1 is that it can return information about the colors in the trees in an approximate solution. That is, our algorithm can maintain a partition of the colors used in X into $I_1, \dots, I_r \subseteq \{1, \dots, k\}$, so that the sum of the costs of the minimum-cost Steiner trees for sets $\bigcup_{i \in I_j} C_i$ is an $(\alpha_2+\varepsilon)$-approximation of SFP. It is worth noting that in estimating the optimal cost, our algorithm does use Steiner points. This means that the MST costs for sets $\bigcup_{i \in I_j} C_i$ of the aforementioned partition may be by an $O(1)$ factor larger than the estimate of the algorithm.

Comparision to a simple exponential-time approach. As we shall discuss in Section 1.2, a simple brute force enumeration combined with linear sketching techniques yields a streaming algorithm also with near-optimal space, but significantly worse running time that is *exponential* in k. Technically, while this approach demonstrates the amazing power of linear sketching, its core is exhaustive search rather than an algorithmic insight, and thus it is quite limited, offering no path for improvements or extensions. Furthermore, the $\mathrm{poly}(k)$ running time in Theorem 1.1 is exponentially better than the exhaustive search, which seems to be a limit of what linear sketching could possibly achieve. Therefore even though the primary focus of streaming algorithms is on their space complexity, the improvement of the running time is critical in terms of pursuing efficient algorithms and making our techniques broadly applicable. Indeed, similar exponential improvements of running time have been of key importance in the advances of various other fundamental streaming problems, for instance, for moment estimation the query time was improved from $\mathrm{poly}(\varepsilon^{-1})$ to $\mathrm{poly}\log(\varepsilon^{-1})$ [KNPW11], and for heavy hitters from $\mathrm{poly}(n)$ to $\mathrm{poly}\log(n)$ [LNNT19].

1.1.1 Technical Contribution: Adapting Arora's Framework to Streaming

We introduce a method to efficiently implement an offline Arora-style [Aro98] dynamic-programming framework based on the quad-tree decomposition in the streaming setting. This method, which is probably the first of its kind for geometric streams, is our main technical contribution.

In the offline setting, Borradaile, Klein, and Mathieu [BKM15] and then Bateni and Hajiaghayi [BH12] extended the Arora's approach to obtain a polynomial-time approximation scheme (PTAS) for SFP. The key insight of these works is that one can tweak the optimal solution so that its cost remains nearly optimal, but it satisfies certain structural properties that allow for designing a suitable dynamic program. In Section 2, we review the structural theorem and the dynamic-programming approach for SFP from [BH12, BKM15] in more detail.

The main difficulty of using the Arora-style approach in low-space streaming is that in general, such approach requires access to all input points, that is, $\Omega(n)$ space to store $\Omega(n)$ leaves at the bottom of the quad-tree input decomposition that have to be considered as basic subproblems. In order to ensure a low-space implementation of the Arora-style framework in the streaming setting, we will use only $O(k \log \Delta)$ non-uniform leaf nodes of the quad-tree, each corresponding to a square. The definition of these leaf nodes is one of the novel ideas needed to make the dynamic-programming approach work in the streaming setting. Moreover, since each internal node in the quad-tree has degree 4, the total number of quad-tree squares to consider is thus $O(k \cdot \log \Delta)$.

The next challenge is that for the dynamic program to run, we need to find an $(\alpha_2 + \varepsilon)$-approximate estimation for each new leaf and each dynamic-programming subproblem associated with it. The definition of leaf squares will enable us to reduce it to estimating the MST cost for a certain subset of points inside the square. It would then be natural to just employ the MST sketch designed in [FIS08] to estimate the MST cost, in a black box manner. However, the leaf squares are not known in advance as we can only find them after processing the stream and thus, it is impossible to build the MST sketch for each leaf square and each subproblem associated with it. To overcome this, we observe that in essence, the MST sketch consists of uniformly sampled points (with suitably rounded coordinates). We thus obtain the MST sketch for each color separately and we only use the sampled points that are relevant for the subproblem to estimate the MST cost for the subproblem, in a way similar to [FIS08].

However, due to restricting the attention to a single subproblem, the original analysis of the MST sketch in [FIS08] has to be modified as we need to deal with additional technical challenges. For instance, we may not sample any point relevant to a leaf square in case there are relatively few points in it. We will need to account for the error arising from this case in a global way, by observing that in such a case, the MST cost inside the leaf square is a small fraction of the overall cost.

Further, to be able to accurately enumerate the subproblems for a leaf square, we need to know the set of color classes that intersect every leaf square, but unfortunately doing so exactly is impossible in the streaming setting. To this end, we employ a δ-net for a small-enough δ, so that the intersection test can be approximately done by only looking at the nearby net points. We show that this only introduces a small error for SFP, and that this δ-net can be constructed in a dynamic stream, using space by only a factor of $\mathrm{poly}\log(\Delta)$ larger than the net. Finally, we apply the dynamic program using our leaf nodes as basic subproblems to obtain the estimation.

1.2 Could Other Approaches Work?

A simple exponential-time streaming algorithm based on linear sketching. An obvious challenge in solving SFP is to determine the connected components of an optimal (or approximate) solution. Each color class must be connected, hence the crucial information is which *colors* are connected together (even though they do not have to be). Suppose momentarily that the algorithm receives an advice with this information, which can be represented as a partition of the color set $[k] = P_1 \sqcup \cdots \sqcup P_l$. Then a straightforward approach for SFP is to solve the Steiner tree problem separately on each part P_j (i.e., the union of some color classes), and report their total cost. In our streaming model, we could apply the aforementioned MST-based algorithm [FIS08], using space $\mathrm{poly}(\varepsilon^{-1}\log\Delta)$, to achieve $(\alpha_2+\varepsilon)$-approximation, and we would need $l \le k$ parallel executions of it (one for each P_j). An algorithm can bypass having such an advice by enumeration, i.e., by trying in parallel all the k^k partitions of $[k]$ and reporting the minimum of all their outcomes. This would still achieve $(\alpha_2+\varepsilon)$-approximation, because each possible partition gives rise to a feasible SFP solution (in fact, this algorithm optimizes the sum-of-MST objective). However, this naive enumeration increases the space and time complexities by a factor of $O(k^k)$. We can drastically improve the space complexity by the powerful fact that the MST algorithm of [FIS08] is based on a *linear sketch*, i.e., its memory contents is obtained by applying a (randomized) linear mapping to the input X. The huge advantage is that linear sketches of several point sets are mergeable. In our context, one can compute a linear sketch for each color class C_i, and then obtain a sketch for the union of some color classes, say some P_j, by simply adding up their linear sketches. These sketches are randomized, and hence, one has to make sure they use the same random coins (same linear mapping), and also to amplify the success probability of the sketches so as to withstand a union bound over all 2^k subsets $P_j \subset [k]$. This technique improves the space complexity and update time to be polynomial in k, basically $\mathrm{poly}(k\varepsilon^{-1}\log\Delta)$, however the query time is still *exponential* in k (see Theorem 5.1 for details).

Tree embedding. Indyk [Ind04] incorporated the low-distortion tree embedding approach of Bartal [Bar96] to obtain dynamic streaming algorithms with $O(\log\Delta)$ ratio for several geometric problems. This technique can be easily applied to SFP as well, but the approximation ratio is $O(\log\Delta)$ which is far from optimal, far from what we are aiming at.

Other $O(1)$-approximate offline approaches. In the regime of $O(1)$-approximation, SFP has been extensively studied using various other techniques, not only dynamic programming. For example, in the offline setting there are several 2-approximation algorithms for SFP using the primal-dual approach and linear programming relaxations [AKR95, GW95, Jai01], and there is also a combinatorial (greedy-type) constant-factor algorithm called *gluttonous* [GK15]. Both of these approaches work in the general metric setting. While there are no known methods to turn the LP approach into low-space streaming algorithms, the gluttonous algorithm of [GK15] might seem amenable to streaming. Indeed, it works similarly to Kruskal's MST algorithm as it also builds components by considering edges in the sorted order by length, and the MST cost estimation in [FIS08] is similar in flavor to Kruskal's algorithm. However, a crucial difference is that the gluttonous algorithm stops growing a component once all terminals inside the component are satisfied, i.e., for each color i, the component either contains all points of C_i, or no point from C_i. This creates a difficulty that the algorithm must know for each component whether or not it is "active" (i.e., not satisfied), and there are up to n components, requiring overall $\Omega(n)$ bits of space. This information is crucial because "inactive" components do not have to be connected to anything else, but they may help to connect two still "active" components in a much cheaper way than by connecting them directly. Apart from these implementation challenges, we have a simple one-dimensional example showing that the approximation ratio of the gluttonous algorithm cannot be better than 2 (moreover, its approximation guarantee in [GK15] is significantly larger than 2). In comparison, our dynamic-programming approach gives a substantially better ratio of $\alpha_2 + \varepsilon$. Nevertheless, it is an interesting open question whether the gluttonous algorithm admits a low-space streaming implementation.

1.3 Related Work

SFP has been extensively studied in operations research and algorithmic communities for several decades. This problem has been also frequently considered as a part of a more general network design problem (see, e.g., [AKR95, GW95, Jai01, MW95]), where one could require for some subsets of vertices to maintain some higher inter-connectivity.

In the classical, offline setting, it is known that the Steiner tree problem, and thus also SFP which is more general, is NP-hard and APX-hard in general graphs and in high-dimensional Euclidean spaces. In general graphs, a 2-approximation algorithm is known due to Agrawal, Klein and Ravi [AKR95] (see also [GW95, Jai01]). These 2-approximation algorithms rely on linear programming relaxations, and the only two combinatorial constant-factor approximations for SFP were recently devised by Gupta and Kumar [GK15] and by Groß et al. [GGK+18]. For low-dimensional Euclidean space, which is the main focus of our paper, Borradaile, Klein, and Mathieu [BKM15] and then Bateni and Hajiaghayi [BH12] obtained $(1 + \varepsilon)$-approximation, i.e., a PTAS, by applying dynamic programming and geometric space decomposition, significantly extending the approach of Arora [Aro98]. Further extensions of the dynamic-programming approach have led to a PTAS for metrics of bounded doubling dimension [CHJ18] and for planar graphs and graphs of bounded treewidth [BHM11].

There has been also extensive work for geometric optimization problems in the dynamic (turnstile) streaming setting, with low space. Indyk [Ind04] introduced this framework and designed $O(\log \Delta)$-estimation algorithms for several basic problems, like MST and matching. Follow-up papers presented a number of streaming algorithms achieving approximation ratio of $1 + \varepsilon$ or $O(1)$ to the cost of Euclidean MST [FIS08], various clustering problems [FS05, HM04], geometric facility location [CLMS13, LS08], earth-mover distance [ABIW09, Ind04], and various geometric primitives (see, e.g., [AN12, Cha06, Cha16, FKZ05]). Some papers have studied geometric problems with superlogarithmic but still sublinear space and in the multipass setting (see, e.g., [ANOY14]). We are not aware of prior results for the (Euclidean) Steiner tree problem nor SFP in the streaming context, although $(1 + \varepsilon)$-approximation of the MST cost [FIS08] immediately gives a $(\alpha_2 + \varepsilon)$-approximation of the Euclidean Steiner tree.

1.4 Future Directions

We believe that our paper is an important step towards understanding the applicability of Arora's framework for low-space streaming algorithms for geometric optimization problems. Still, our work leaves a number of open problems which should be of broad interest for streaming researchers. We refer the reader to Section 6 for an exhaustive list of open problems related to our work, but here we mention the main open problem: Our approximation ratio $\alpha_2 + \varepsilon$ matches the current state of the art approximation ratio for the Steiner tree problem in geometric streams. Hence, any improvement to our approximation ratio would require to first improve the approximation for Steiner tree, even in insertion-only streams. This naturally leads to the main open problem of obtaining a $(1 + \varepsilon)$-approximation for Steiner tree in geometric streams using only $\mathrm{poly}(\varepsilon^{-1} \log \Delta)$

space.

2 Preliminaries

2.1 Notations

For $x, y \in \mathbb{R}^2$, let $\mathrm{dist}(x, y) := \|x-y\|_2$. For two subsets $S, T \in \mathbb{R}^2$, let $\mathrm{dist}(S, T) := \min_{x\in S, y\in T} \mathrm{dist}(x, y)$. For $S \subset \mathbb{R}^2$, let $\mathrm{diam}(S) := \max_{x,y\in S} \mathrm{dist}(x, y)$. A *$\rho$-packing* $S \subset \mathbb{R}^2$ is a point set such that $\forall x, y \in S$, $\mathrm{dist}(x, y) \geq \rho$. A *$\rho$-covering* of X is a subset $S \subset \mathbb{R}^2$, such that $\forall x \in X$, $\exists y \in S$, $\mathrm{dist}(x, y) \leq \rho$. We call $S \subset \mathbb{R}^2$ a *ρ-net* for X if it is both a ρ-packing and a ρ-covering for X.

Fact 2.1 (Packing Property, cf. [Pol90, Lemma 4.1])**.** *A ρ-packing $S \subset \mathbb{R}^d$ has size $|S| \leq \left(\frac{3\,\mathrm{diam}(S)}{\rho}\right)^d$.*

Metric graphs. We call a weighted undirected graph $G = (X, E, w)$ a *metric graph* if for every edge $\{u, v\} \in E$, $w(u, v) = \mathrm{dist}(u, v)$, and we let $w(G)$ to be the sum of the weights of edges in G. A solution F of SFP may be interpreted as a metric graph. For a set of points S (e.g., S can be a square), let $F|_S$ be the subgraph of F formed by edges whose both endpoints belong to S. Note that we think of F as a *continuous* graph in which every point of an edge is itself a vertex, so $F|_S$ may be interpreted as a geometric intersection of F and S.

Randomly-shifted quad-trees [Aro98]. Without loss of generality, suppose that Δ is a power of 2, and let $L := 2\Delta$. A quad-tree sub-division is constructed on $[L]^2$. In the quad-tree, each node u corresponds to a square R_u and if it's not a leaf, it has four children, whose squares partition R_u. The squares in the quad-tree are of side-lengths that are powers of 2, and we say a square R is of level i if its side-length is 2^i (this is also the level of its corresponding node in the quad-tree, where leaves have level 0 and the root is at level $\log_2 L$). The whole quad-tree is shifted by a random vector in $[-\Delta, 0]^2$. Throughout, we assume a randomly-shifted quad-tree has been sampled from the very beginning. When we talk about a quad-tree square R, we interpret it as the point set that consists of both the boundary and the internal points. For $i = 0, \ldots \log_2 L$, let *2^i-grid* $\mathcal{G}_i \subset \mathbb{R}^2$ be the set of centers of all level-i squares in the quad-tree.

2.2 Review of Dynamic Programming (DP) [BH12, BKM15]

The PTAS for geometric SFP in the offline setting [BH12, BKM15] is based on the quad-tree sub-division framework of Arora [Aro98], with modifications tailored to SFP. For each square R in the (randomly-shifted) quad-tree,

- $O(\varepsilon^{-1} \log L)$ equally-spaced points on the four boundary edges are designated as *portals*; and
- the $\gamma \times \gamma$ sub-squares of R are designated as *cells* of R, denoted $\mathsf{cell}(R)$, where $\gamma = \Theta(\varepsilon^{-1})$ is a power of 2.

For each square R in the quad-tree, let ∂R be the boundary of R (which consists of four segments). The following is the main structural theorem from [BH12], and an illustration of it can be found in Figure 1a.

Theorem 2.2 ([BH12])**.** *For an optimal solution F of SFP, there is a solution F' (defined with respect to the randomly-shifted quad-tree), such that*

1. *$w(F') \leq (1 + O(\varepsilon)) \cdot w(F)$ with constant probability (over the randomness of the quad-tree);*

2. *For each quad-tree square R, $F'|_{\partial R}$ has at most $O(\varepsilon^{-1})$ components, and each component of $F'|_{\partial R}$ contains a* portal *of R;*

3. *For each quad-tree square R and each cell P of R, if two points $x_1, x_2 \in X \cap P$ are connected to ∂R via F', then they are connected in $F'|_R$; this is called the* cell property*.*

It suffices to find the optimal solution that satisfies the structure defined in Theorem 2.2. This is implemented using dynamic programming (DP), where a subproblem of the DP is identified as a tuple (R, A, f, Π), specified as follows:

- R is a quad-tree square;
- A is a set of at most $O(\varepsilon^{-1})$ active portals through which the local solution enters/exits R;
- $f : \mathsf{cell}(R) \to 2^A$ s.t. for $S \in \mathsf{cell}(R)$, $f(S)$ represents the subset of A that S connects to;
- Π is a partition of A, where active portals in each part of Π have to be connected outside of R (in a larger subproblem).

The use of R and A is immediate, and f is used to capture the connectivity between cells and portals (this suffices because we have the "cell property" in Theorem 2.2). Finally, Π is used to ensure feasibility, since a global connected component may be broken into several components in square R, and it is important to record whether or not these components still need to be connected from outside of R. An optimal solution for subproblem (R, A, f, Π) is defined as a minimum weight metric graph in R that satisfies the constraints A, f, Π.

Remark 2.3. Strictly speaking, we use a simplified definition of DP subproblems, compared to [BH12]. Namely, one can additionally require that for any two cells $S, S' \in \mathsf{cell}(R)$, either $f(S) = f(S')$ or $f(S) \cap f(S') = \emptyset$ and that any active portal in A appears in $f(S)$ for some cell S. Then, f defines a partition of $\mathsf{cell}(R)$ and of A into local components inside R (taking into account only components connected to ∂R), and Π should encode which local components need be connected from the outside of R, implying that Π should be a partition of local components (instead of A). Thus, Π can also be thought of as a partition of the partition of A induced by f. We chose to give a more relaxed definition of DP subproblems as it is sufficient for describing how to implement the DP approach in the streaming setting.

Standard combinatorial bounds show that the number of subproblems associated with each square is bounded by $(\varepsilon^{-1} \cdot \log \Delta)^{O(\varepsilon^{-2})}$ (see [BH12]).

3 Streaming Dynamic Programming: k^3-time-and-space Algorithm

In this section, we prove our main result, Theorem 1.1, restated with more precise bounds. Formally, we call the time for processing inserting/deleting one point as *update time*, and for reporting the estimate of OPT the *query time*.

Theorem 3.1. *For any integers $k, \Delta \geq 1$ and any $0 < \varepsilon < 1/2$, one can with high probability $(\alpha_2 + \varepsilon)$-approximate the SFP cost of an input $X \subseteq [\Delta]^2$ presented as a dynamic geometric stream, using space and update time of $k^3 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \cdot \log \Delta)$ and with query time bounded by $k^3 \cdot \mathrm{poly}(\log k) \cdot (\varepsilon^{-1} \cdot \log \Delta)^{O(\varepsilon^{-2})}$.*

Overview. As reviewed in Section 2.2, a PTAS has been shown for SFP in the offline setting [BH12, BKM15]. Our overall approach for the streaming algorithm is to modify this algorithm, which is based on dynamic programming (DP). Observe that one important reason that the DP requires $\Omega(n)$ space is that $\Omega(n)$ leaves in the quad-tree have to be considered as basic subproblems which correspond to singletons. To make the DP use only $\widetilde{O}(\text{poly}(k))$ space, we wish to use only $\widetilde{O}(\text{poly}(k))$ leaf nodes. Indeed, since each internal node in the quad-tree has degree 4, the total number of squares to consider is thus $\widetilde{O}(\text{poly}(k))$. Furthermore, we design an algorithm that runs in time and space $\widetilde{O}(\text{poly}(k))$ and finds an $(\alpha_2 + \varepsilon)$-approximate estimation for each new leaf and each DP subproblem associated with it. Finally, we apply the DP using such leaves as basic subproblems to obtain the estimation. We start with a description of this approach in the offline setting (Section 3.1), and we make it streaming in Section 3.2. We then give the proof of Theorem 3.1 in Section 3.3.

3.1 Offline Algorithm

New definition of basic subproblems. Each of our new leaves in the DP will be a *simple square* defined below. The idea behind the definition is also simple: If no color is contained in R, then all points inside R must be connected to ∂R, so we can make better use of the cell property in Theorem 2.2.

Definition 3.2 (Simple squares)**.** We call a square R *simple* if for every $1 \leq i \leq k$, $C_i \cap R \neq C_i$. In other words, there is no color totally contained in R.

We note that the number of all possible simple squares can still be large (in particular, any empty square is simple as well as any square containing a single point of color C_i with $|C_i| \geq 2$), and we use Lemma 3.3 below to show the existence of a small subset of simple squares that covers the whole instance and can be found efficiently. Our new leaves are naturally defined using such subset of squares.

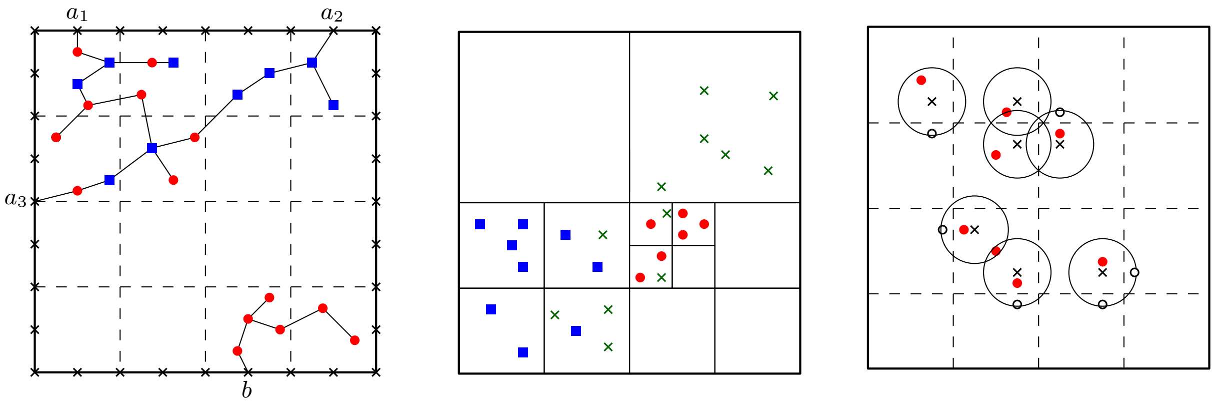
Lemma 3.3. *There is a subset $\mathcal{R}$ of disjoint simple squares, such that the union of the squares in $\mathcal{R}$ covers X, and $|\mathcal{R}| = O(k \cdot \log \Delta)$.*

Proof. Consider the recursive procedure specified in Algorithm 1 that takes as input a square R and returns a set of disjoint simple squares $\mathcal{R}$ that covers R; see Figure 1b for an illustration of the outcome of the procedure. For our proof, we apply the procedure with R being the root square covering the whole instance.

Algorithm 1 Algorithm for finding simple squares

```
procedure SIMP-SQUARE($R$)
    if $R$ is simple then
        return $\{R\}$
    else
        let $\{R_i\}_i$ be the child squares of $R$ in the quad-tree
        return $\bigcup_i$ SIMP-SQUARE($R_i$)
    end if
end procedure
```

Suppose the procedure returns $\mathcal{R}$. We call a square R intermediate square if it is a square visited in the execution of the algorithm and it is not simple (i.e., R contains a color class). We observe that $|\mathcal{R}|$ is $O(1)$ times the number of intermediate squares. On the other hand, each color C_i can be totally contained in at most $O(\log \Delta)$ intermediate squares. Therefore, $|\mathcal{R}| = O(k \cdot \log \Delta)$. □



(a) structural property

(b) simple squares

(c) compatibility checking

Figure 1: Illustrations of the structural properties of Theorem 2.2 (Figure 1a), construction of simple squares by Algorithm 1 (Figure 1b) and the approximate compatibility checking idea in Section 3.2.2 (Figure 1c). Figure 1a shows a square R with portals (crosses) on ∂R, the 4×4 cells of R, and the part of solution $F'|_R$, such that $F'|_R$ passes ∂R through four portals a_1, a_2, a_3, b on the sides, and in each cell, points that are connected by F' to ∂R in the cell are connected in R. Figure 1b demonstrates the 13 simple squares constructed by Algorithm 1 for the three colors (noting that the 5 empty squares are also included as simple squares). In Figure 1c, red points are data points, cross points are the net points constructed from the data, and the black hollowed points are the added points for cells that are close-enough to a net point (for simplicity, not shown for cells containing a data point).

Approximation algorithm for subproblems on simple squares. Fix some simple square R. We now describe how each DP subproblem (R, A, f, Π) associated with R can be solved directly using an α_2-approximation algorithm that is amenable to the streaming setting.

Since R is a simple square, every point in R has to be connected to the outside of R, as otherwise the color connectivity constraint is violated. Hence, by the cell property of Theorem 2.2, for every cell $R' \in \mathsf{cell}(R)$, all points in R' are connected in R.

Therefore, we enumerate all possible partitions of $\mathsf{cell}(R)$ that is consistent with the f constraint. For each partition, we further check whether it satisfies the constraint defined by Π. To do so, for each cell $R' \in \mathsf{cell}(R)$, we scan through all colors, and record the set of colors $\mathcal{C}_{R'} \subseteq \mathcal{C}$ that intersects R'. The $\mathcal{C}_{R'}$'s combined with the f constraint as well as the enumerated connectivity between cells suffice for checking the Π constraint.

Observe that every feasible solution of the subproblem corresponds to the above-mentioned partition of cells. Therefore, to evaluate the cost of the subproblem, we evaluate the sum of the MST costs of the parts in each partition and return the minimum one.

The time complexity for evaluating each subproblem is bounded since $|A| = O(\varepsilon^{-1})$ and $|\mathsf{cell}(R)| = O(\varepsilon^{-2})$. The approximation ratio is α_2 because we use MST instead of Steiner tree for evaluating the cost. Using MST will enable us to implement this algorithm in the streaming setting.

3.2 Building Blocks for Streaming Algorithm

We implement the offline algorithm in the streaming setting. As we shall see, the offline algorithm consists of several important steps that are nontrivial to implement in the streaming setting. Thus, we start with presenting the streaming building blocks of these important steps, in Sections 3.2.1 to 3.2.3.

3.2.1 Constructing Simple Squares in the Streaming Setting

The first step is to construct a set of simple squares, as in Lemma 3.3, and an offline construction is outlined in Algorithm 1. For the streaming construction of simple squares, we observe that the key component of Algorithm 1 is a subroutine that tests whether a given square is simple or not. To implement the subroutine, we use a streaming algorithm to compute the bounding square for each color, and we test whether a given square contains any bounding square as a sub-square.

Lemma 3.4. *Algorithm 1 can be implemented in the streaming setting, using space* $O(k \operatorname{poly}\log \Delta)$ *and in time* $O(k \operatorname{poly}\log \Delta)$ *per stream update, with success probability at least* $1 - \operatorname{poly}(\Delta^{-1})$.

Proof. We show that for a point set $S \subseteq [\Delta]^2$ and a quad-tree decomposition, the smallest quad-tree square that contains S as a subset can be computed in the dynamic streaming setting, as stated in Algorithm 2. We argue that Algorithm 2 indeed finds the minimal enclosing quad-tree square

Algorithm 2 Finding the minimal enclosing quad-tree square

```
procedure Min-Sqaure($S \subseteq [\Delta]^2$)
    for each $1 \le i \le \log_2 L$, maintain sketch $\mathcal{K}_i$ of Lemma A.1 for the $2^i$-grid $\mathcal{G}_i$ (see Section 2), with parameter $T = 1$
    for insertion/deletion of point $x$ in the stream do
        for $i \leftarrow 0, \ldots, \log_2 L$ do
            let $y \in \mathcal{G}_i$ be the grid point for the level-$i$ square that $x$ belongs to
                ▷ recall that grid points in $\mathcal{G}_i$ are centers of level-$i$ squares
            increase/decrease the frequency of $y$ by 1 in sketch $\mathcal{K}_i$
        end for
    end for    ▷ the stream terminates, and the computing phase starts
    find the smallest $i$ such that sketch $\mathcal{K}_i$ returns exactly one element, and return the corresponding square
end procedure
```

for a point set S, with high probability (w.h.p.). Suppose R is the minimal quad-tree square that contains S, and suppose R is of level i. Then all points in S correspond to the same grid point y in $\mathcal{G}_i$, so $\mathcal{K}_i$ returns exactly square R w.h.p. On the other hand, for any level $i' < i$, sketch $\mathcal{K}_{i'}$ either reports that the number of non-empty grid squares is larger than 2, or returns 2 squares w.h.p. This concludes that Algorithm 2 finds exactly the minimal enclosing square of S w.h.p.

We apply Algorithm 2 for each color $C_i \in \mathcal{C}$. After the stream ends, we get the bounding squares $\{R_i\}_i$ for colors $\{C_i\}_i$. Then we execute Algorithm 1 using the bounding squares $\{R_i\}$. In particular, to test whether a square R is simple or not, it suffices to scan through all the bounding squares $\{R_i\}_i$, and R is simple if and only if R does not contain any R_i as a sub-square (note that if $R = R_i$ for some i, then R is not simple).

By the guarantee of Lemma A.1, the space is bounded by $O(k \operatorname{poly}\log \Delta)$, and the overall success probability is at least $1 - \operatorname{poly}(\Delta^{-1})$. □

3.2.2 Approximate Compatibility Checking

Suppose we apply Lemma 3.4, and it yields a set of simple squares $\mathcal{R}$. We proceed to evaluate the cost of the DP subproblems associated with each simple square. Fix a simple square $R \in \mathcal{R}$.

We next describe how to evaluate the cost for every subproblem associated with R, in a streaming manner.

Suppose we are to evaluate the cost of a subproblem (R, A, f, Π). Since R is known, we have access to $\mathsf{cell}(R)$, and hence, we can enumerate the connectivity between the cells, which is a partition of $\mathsf{cell}(R)$, on-the-fly without maintaining other information about the input. Similarly, we can check the compatibility of the partition of cells with the f constraint, since the constraint only concerns the information about A and the partition.

Then, when we check the compatibility of the partition of cells with Π, in the offline setting we need to compute the set of colors $\mathcal{C}_{R'} \subseteq \mathcal{C}$ that a cell R' intersects. However, we note that computing this set $\mathcal{C}_{R'}$ is difficult in the streaming setting, even if there is only one color C. Indeed, testing whether color C has an intersection with cell R' can be immediately reduced to the INDEX problem (see e.g. [KN97] or Section 4 for the definition), which implies an $\Omega(n)$ space bound, where n is the number of points of color C. Therefore, we need to modify the offline algorithm, and only test the intersection approximately.

To implement the approximate testing, for every color $C \in \mathcal{C}$, we impose a $\delta \cdot \mathrm{diam}(C)$-net N_C for C (as defined in Section 2), where $\delta := O\left(\varepsilon^3 (k \log \Delta)^{-1}\right)$. We show that such a net can be constructed in the streaming setting in Lemma 3.5. To be exact, the streaming algorithm in Lemma 3.5 returns a set of net points N_C such that for any point $x \in N_C$ at least one point in C is within distance $\delta \cdot \mathrm{diam}(C)$ from x (so N_C does not contain net points that are far away from any point in C). Hence, take $D_C := \mathrm{diam}(N_C)$, and we have $D_C \in (1 \pm \delta) \cdot \mathrm{diam}(C)$. Then, for each cell $R' \in \mathsf{cell}(R)$ of each simple square R, we examine each point in N_C, and if $\mathrm{dist}(R', N_C) \le \delta \cdot D_C$, we add a new point $x \in R'$ such that $\mathrm{dist}(x, N_C) \le \delta \cdot D_C$ to the stream, and assign it color C. Furthermore, we declare C intersects R'. This idea is visually demonstrated in Figure 1c.

Lemma 3.5. *There is an algorithm that for every $0 < \rho \le 1$ and every point set $S \subset \mathbb{R}^2$ provided as a dynamic geometric stream, computes a subset $N_S \subset \mathbb{R}^2$ that is a $\rho \cdot \mathrm{diam}(S)$-net for S such that for every $x \in N_S$ there exists $y \in S$ with $\mathrm{dist}(x, y) \le \rho \cdot \mathrm{diam}(S)$, with probability at least $1 - \mathrm{poly}(\Delta^{-1})$, using space $O(\rho)^{-2} \cdot \mathrm{poly}\log \Delta$, and running in time $O(\rho)^{-2} \cdot \mathrm{poly}\log \Delta$ per stream update.*

Proof. We give the procedure in Algorithm 3. It makes use of Lemma A.1 in a way similar to Algorithm 2. The space and time complexity as well as the failure probability follow immediately

Algorithm 3 Streaming algorithm for constructing the net

```
procedure Net($S \subseteq [\Delta]^2, \rho \in (0, 1]$)
    for each $1 \le i \le \log_2 L$, maintain sketch $\mathcal{K}_i$ of Lemma A.1 for the $2^i$-grid $\mathcal{G}_i$, with parameter $T = (3\rho^{-1})^2$
    for insertion/deletion of point $x$ in the stream do
        for $i \leftarrow 0, \ldots, \log_2 L$ do
            let $y \in \mathcal{G}_i$ be the grid point for the level-$i$ square that $x$ belongs to
            increase/decrease the frequency of $y$ by 1 in sketch $\mathcal{K}_i$
        end for
    end for                    ▷ the stream terminates, and the computing phase starts
    find the smallest $i$ such that sketch $\mathcal{K}_i$ reports the number of elements is $\le 2T$
    return the $\le 2T$ grid points in $\mathcal{G}_i$ that $\mathcal{K}_i$ reports
end procedure
```

from the guarantee of Lemma A.1. We now analyze the correctness. Observe that by Fact 2.1, every $\rho \cdot \mathrm{diam}(S)$-net for S has size at most $(3\rho^{-1})^2 = T$, so the point set returned by the sketch is a net that can only be finer. Also, by the construction, if a point y is reported by Algorithm 3, then there must be a point $x \in S$ such that x is inside the square whose center is y. Therefore, for every $y \in N_C$, there is $x \in S$ such that $d(x, y) \leq \rho \cdot \mathrm{diam}(S)$. This finishes the proof of Lemma 3.5. □

In fact, such procedure of adding points is oblivious to the subproblem, and should be done only once as a *pre-processing step* before evaluating any subproblems. Therefore, the subproblems are actually evaluated on a new instance $(X', \mathcal{C}')$ after the pre-processing. Since we apply Lemma 3.5 for every color i, and by the choice of δ, the space complexity for the pre-processing step is $O\left(k^3 \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta)\right)$, and the time complexity per update is bounded by this quantity. Next, we show that the error introduced by the new instance is well bounded.

Lemma 3.6. *Let* OPT *be the optimal SFP solution for the original instance* $(X, \mathcal{C})$*, and let* OPT' *be that for* $(X', \mathcal{C}')$*. Then* $w(\mathrm{OPT}) \leq w(\mathrm{OPT}') \leq (1 + \varepsilon) \cdot w(\mathrm{OPT})$*.*

Proof. Since OPT' is a feasible solution for $(X, \mathcal{C})$, we obtain $w(\mathrm{OPT}) \leq w(\mathrm{OPT}')$ by the optimality of OPT. It remains to prove the other side of the inequality.

Recall that for every color C, we use Lemma 3.5 to obtain a $\delta \cdot \mathrm{diam}(C)$-net N_C and estimate $\mathrm{diam}(C)$ using $D_C := \mathrm{diam}(N_C)$. Then, for every cell R' of every simple square, if $\mathrm{dist}(R', N_C) \leq \delta \cdot D_C$ for some color C (chosen arbitrarily if there are more), we add a point x to color class C satisfying $d(x, N_C) \leq \delta \cdot D_C$. Moreover, for any other color $C' \neq C$ with $\mathrm{dist}(R', N_{C'}) \leq \delta \cdot D_{C'}$, we add the same point x to color class C'. Note that we only add at most one distinct point for each cell. Let z be a point $z \in N_C$ with $d(x, z) \leq \delta \cdot D_C$. Adding point x increases OPT by at most $2\delta \cdot D_C \leq 3\delta \cdot \mathrm{diam}(C)$, since one can connect x to a point y in C such that $\mathrm{dist}(y, z) \leq \delta \cdot \mathrm{diam}(C)$ (the existence of y is guaranteed by Lemma 3.5).

Since there are in total at most $O(k \log \Delta \cdot \varepsilon^{-2})$ cells in all simple squares by Lemma 3.3, the total increase of the cost is at most

$$O(\delta \cdot k \log \Delta \cdot \varepsilon^{-2}) \cdot \max_{C \in \mathcal{C}} \mathrm{diam}(C) \leq \varepsilon \max_{C \in \mathcal{C}} \mathrm{diam}(C) \leq \varepsilon w(\mathrm{OPT}), \tag{1}$$

using the definition of δ and $w(\mathrm{OPT}) \geq \max_{C \in \mathcal{C}}(\mathrm{diam}(C))$. We conclude that $w(\mathrm{OPT}') \leq (1 + \varepsilon) \cdot w(\mathrm{OPT})$. □

3.2.3 Evaluating Basic Subproblems in the Streaming Setting

After we obtain the new instance $(X', \mathcal{C}')$, we evaluate the cost for every subproblem (R, A, f, Π). Because of the modification of the instance, we know for sure the subset of colors $\mathcal{C}_{R'}$ for each cell R'. To evaluate the subproblem, recall that we start with enumerating a partition of $\mathsf{cell}(R)$ that is compatible with the subproblem, which can be tested efficiently using $\mathcal{C}_{R'}$'s. Suppose now $\{P_i := R_i \cup A_i\}_{i=1}^t$ is a partition of $\mathsf{cell}(R) \cup A$ that we enumerated (recalling that A is the set of active portals, which needs to be connected to cells in a way that is compatible to the constraint f). Then, as in the offline algorithm, we evaluate $\mathrm{MST}(P_i)$ of each part P_i, and compute the sum of them, i.e. $\sum_{i=1}^t \mathrm{MST}(P_i)$, however, we need to show how to do this in the streaming setting.

Frahling et al. [FIS08] designed an algorithm that reports a $(1+\varepsilon)$-approximation for the value of the MST of a point set presented in a dynamic stream, using space $O(\varepsilon^{-1} \log \Delta)^{O(1)}$. Furthermore, as noted in Section 1, their algorithm maintains a linear sketch. Now, a natural idea is to apply this MST sketch, that is, create an MST sketch for each color, which only takes $k \cdot O(\varepsilon^{-1} \log \Delta)^{O(1)}$ space. Then, for each $P_i = R_i \cup A_i$, we compute the set of intersecting colors, and we create a new MST sketch $\mathcal{K}$ by first adding up the MST sketches of these colors (recalling that they are linear sketches), and then adding the active portals connected to P_i to the sketch. We wish to query the sketch $\mathcal{K}$ for the cost of $\mathrm{MST}(P_i)$.

However, this idea cannot directly work, since the algorithm by [FIS08] only gives the MST value for *all* points represented by $\mathcal{K}$, instead of the MST value for a subset P_i. Therefore, we will modify the MST sketch to answer the value of the MST on a subset of points of interest.

Brief review of the MST sketch. We give a brief overview of the algorithm of [FIS08] before we explain how we modify it. The first observation (already from [CRT05]) is that the MST cost can be written as a weighted sum of the number of connected components in metric threshold graphs, which are obtained from the complete metric graph of the point set by removing edges of length larger than a threshold τ. Essentially, the idea is to count the number of MST edges of length larger than τ.

To estimate the number of components in a threshold graph, we round the points to a suitable grid and sample a small number of rounded points uniformly, using ℓ_0-samplers. An ℓ_0-sampler is a data structure that processes a dynamic stream (possibly containing duplicate items), succeeds with high probability, and conditioned on it succeeding, it returns a random item from the stream such that any item in the stream is chosen with the same probability $1/n$, where n is the ℓ_0 norm of the resulting frequency vector, i.e., the number of distinct items in the stream (see Lemma 3.10 for a more precise statement). For each sampled (rounded) point y, the algorithm in [FIS08] runs a stochastic-stopping BFS from y and in particular, it checks if it explores the whole component of y within a random number of steps. We note that this requires an extended ℓ_0-sampler that also returns the neighboring points for each sampled point, as presented in [FIS08] and stated in Lemma 3.10. The MST cost is estimated by a weighted sum of the number of completed BFS's, summed over all levels.

Generalizing the MST algorithm to handle subset queries. Fix some part P_i. Recall that the P_i's always consist of at most $O(\varepsilon^{-2})$ cells (which are quad-tree squares), plus $O(\varepsilon^{-2})$ active portal points. Hence, a natural first attempt is to make the ℓ_0-samplers to sample only on these clipping squares defined by P_i. Unfortunately, this approach would not work, since the squares are not known in advance and may be very small (i.e., degenerate to a point), so sampling a point from them essentially solves the INDEX problem.

Therefore, when estimating $\mathrm{MST}(P_i)$, we still use the original ℓ_0-samplers, and we employ a careful sampling and estimation step. We sample from the whole point set maintained by the sketch $\mathcal{K}$ by querying the ℓ_0-samplers, but we only keep the sampled points contained in P_i. We execute the stochastic BFS from these points that are kept, restricting the BFS to the points contained in P_i.

One outstanding problem of this sampling method is that if the number of points in P_i, or to be exact, the number of non-zero entries of level-i ℓ_0-samplers, is only a tiny portion of that of the full sketch, then with high probability, we do not sample any point from P_i at all. Hence, in this case, no stochastic BFS can be performed, and we inevitably answer 0 for the number of successful BFS's. This eventually leads to an additive error. We summarize the additive error and the whole idea of the above discussions in Lemma 3.7.

Lemma 3.7. *There is an algorithm that for every $0 < \varepsilon < 1$, integer $k, \Delta \geq 1$, and every set of points $S \subseteq [\Delta]^2$ presented as a dynamic geometric stream, maintains a linear sketch of size $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$. For every query $(R, \{R_j\}_{j=1}^t, A)$ (provided after the stream ends) satisfying*

1. *R is a simple square, A is a subset of portals of R, and*

2. $\{R_j\}_{j=1}^t \subseteq \mathsf{cell}(R)$,

the algorithm computes from the linear sketch a real number E such that with probability at least $1 - \exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta))$,

$$\mathrm{MST}(P) \le E \le (1+\varepsilon) \cdot \mathrm{MST}(P) + O\left(\frac{\mathrm{poly}(\varepsilon)}{k \log \Delta}\right) \cdot \mathrm{MST}(S)\,,$$

where $P = \left(\bigcup_{j=1}^t R_j\right) \cup A$. The algorithm runs in time $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$ per update and the query time is also $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$.

This lemma constitutes the main algorithm for the evaluation of the subproblem. Note that we only need to prove it for one point set S, since the sketch is linear. Indeed, when applying Lemma 3.7, we obtain the sketch for each color separately from the stream, and for every query, we first merge the sketches of colors relevant to the query and add query portals to the resulting sketch. By linearity, this is the same as if we obtain the sketch for all these colors and portals at once. We postpone the proof of Lemma 3.7 to Section 3.4, where we give a more detailed discussion of the technical issues and our novel ideas to overcome them.

3.3 Proof of Theorem 3.1

We first restate Theorem 3.1 for convenience.

Theorem 3.1. *For any integers $k, \Delta \ge 1$ and any $0 < \varepsilon < 1/2$, one can with high probability $(\alpha_2 + \varepsilon)$-approximate the SFP cost of an input $X \subseteq [\Delta]^2$ presented as a dynamic geometric stream, using space and update time of $k^3 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \cdot \log \Delta)$ and with query time bounded by $k^3 \cdot \mathrm{poly}(\log k) \cdot (\varepsilon^{-1} \cdot \log \Delta)^{O(\varepsilon^{-2})}$.*

We combine the above building blocks to prove Theorem 3.1. We start with a description of the complete algorithm (Algorithm 4). The space and update time follow immediately from the description of Algorithm 4 and from Theorem 2.2 and Lemmas 3.4, 3.5 and 3.7.

The query time is bounded by

$$\begin{aligned} O(k \cdot \log \Delta) \cdot &(\varepsilon^{-1} \cdot \log \Delta)^{O(\varepsilon^{-2})} \cdot \varepsilon^{-O(\varepsilon^{-1})} \cdot k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta) \\ &\le k^3 \cdot \mathrm{poly}(\log k) \cdot (\varepsilon^{-1} \log \Delta)^{O(\varepsilon^{-1})} \end{aligned}$$

where $O(k \cdot \log \Delta)$ is the number of simple squares (and thus, up to an $O(1)$ factor, the number of quad-tree squares for which we evaluate DP subproblems), $(\varepsilon^{-1} \cdot \log \Delta)^{O(\varepsilon^{-2})}$ is the number of subproblems associated with each square (see Section 2.2), $\varepsilon^{-O(\varepsilon^{-1})}$ is the number of MST queries evaluated for each subproblem, and each MST query takes $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$ time by Lemma 3.7.

To bound the failure probability, we use a union bound over the failure probabilities of all applications and queries of the streaming algorithms as well as the error bound in Theorem 2.2. We observe that Theorem 2.2 incurs an $O(1)$ failure probability, and every other steps, except for the use of Lemma 3.7, have a failure probability of $\mathrm{poly}(\Delta^{-1})$. Since we have $k \cdot (\varepsilon^{-1} \cdot \log \Delta)^{O(\varepsilon^{-2})}$ basic subproblems (see Section 2), and for each basic subproblem we need to evaluate at most $\varepsilon^{-O(\varepsilon^{-1})}$ MST queries, the total failure probability of evaluating the subproblems is at most

$$k \cdot (\varepsilon^{-1} \cdot \log \Delta)^{O(\varepsilon^{-2})} \cdot \varepsilon^{-O(\varepsilon^{-1})} \cdot \exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta)) \le \mathrm{poly}(\Delta^{-1})\,,$$

by the guarantee of Lemma 3.7. Therefore, we conclude that the failure probability is then at most $\frac{2}{3}$. It remains to analyze the error.

[2] We need to use the same randomness for sketches $\{\mathcal{K}_C^{(3)}\}$ among all colors C so that they can be combined later.

Algorithm 4 Main streaming algorithm

```
1: procedure SFP($X$, $\mathcal{C}$)
2:     initialize a sketch $\mathcal{K}^{(1)}$ of Lemma 3.4, a set of sketches of Lemma 3.5 $\{\mathcal{K}_C^{(2)}\}_{C\in\mathcal{C}}$ for every color $C \in \mathcal{C}$ with parameter $\delta := \mathrm{poly}(\varepsilon)(k \log \Delta)^{-1}$, and a set of (linear) sketches[2]of Lemma 3.7 $\{\mathcal{K}_C^{(3)}\}_{C\in\mathcal{C}}$ for every color $C \in \mathcal{C}$
3:     for every insertion/deletion of point $x$ of color $C$ do
4:         insert/delete point $x$ in sketches $\mathcal{K}^{(1)}, \mathcal{K}_C^{(2)}, \mathcal{K}_C^{(3)}$
5:     end for                                          ▷ the stream terminates
6:     use sketch $\mathcal{K}^{(1)}$ to compute a set of simple squares $\mathcal{R}$          ▷ see Section 3.2.1
7:     for each color $C \in \mathcal{C}$, use sketch $\mathcal{K}_C^{(2)}$ to compute a set of net points $N_C$, and let $D_C := \mathrm{diam}(N_C)$          ▷ $D_C$ is a $(1 \pm \varepsilon)$-approximation for $\mathrm{diam}(C)$
8:     initialize a Boolean list $\mathcal{I}$ that records whether a cell of a simple square and a color intersects          ▷ This uses space at most $O(k \cdot \log \Delta \cdot \mathrm{poly}(\varepsilon^{-1}))$
9:     for every $R \in \mathcal{R}$, $R' \in \mathsf{cell}(R)$ do
10:        if $\mathrm{dist}(N_C, R') \le \rho \cdot D_C$ for some color $C$ then
11:            let $x \in R'$ be a point such that $\mathrm{dist}(x, N_C) \le \rho \cdot D_C$
12:            for every color $C'$ with $\mathrm{dist}(N_{C'}, R') \le \rho \cdot D_{C'}$ do
13:                add $x$ to $\mathcal{K}_{C'}^{(3)}$ and record in $\mathcal{I}$ that $R'$ intersects color $C'$          ▷ see Section 3.2.2
14:            end for
15:        end if
16:    end for
17:    for each simple square $R$ and an associated subproblem $(R, A, f, \Pi)$ do
18:        for each partition of $\mathsf{cell}(R)$ do
19:            if the partition is compatible with the subproblem then          ▷ see Section 3.2.2
20:                for each part $R_j$ in the partition do
21:                    let $A_j \subseteq A$ be the set of active portals that $R_j$ connects to
22:                    create linear sketch $\mathcal{K}'$, by adding up $\mathcal{K}_C^{(3)}$ for every $C$ intersecting a cell in $R_j$          ▷ the intersection information is recorded in $\mathcal{I}$
23:                    add points in $A_j$ to sketch $\mathcal{K}'$
24:                    query sketch $\mathcal{K}'$ for the value of the MST of the part $R_j$ and portals $A_j$ (as in Lemma 3.7)          ▷ see Section 3.2.3
25:                end for
26:                store the sum of the queried values of $\mathrm{MST}(R_j, A_j)$ as the estimated cost for the subproblem
27:            end if
28:        end for
29:    end for
30:    invoke the DP (as in [BH12]) using the values of basic subproblem estimated as above
31:    return the DP value (for the root square with no active portals)
32: end procedure
```

Error analysis. For the remaining part of the analysis, we condition on no failure of the sketches used in Algorithm 4 and on that the error bound in Theorem 2.2 holds. By Lemma 3.6, for the part of evaluating the basic subproblems (Line 17 to Line 29 of Algorithm 4), the actual instance that the linear sketches work on is $(1 + O(\varepsilon))$-approximate. Hence, it suffices to show the DP value is accurate to that instance.

First of all, our estimation is never an underestimate, by Lemma 3.7 and since all partitions that we enumerated are compatible with the subproblems; see Section 3.2.2. Hence, it remains to upper bound the estimation. Consider an optimal DP solution F, which we interpret as a metric graph (see Section 2). Then we create a new solution F' from F by modifying F using the following procedure. For each simple square R, we consider $F|_R$ which is the portion of F that is totally inside of R (see Section 2). For each component $S \subset R$ in $F|_R$, let S' be the point set formed by removing all Steiner points from S, except for portals of R (note that we remove portals of subsquares of R if they appear in S). Then, for each component S, we replace the subtree in F that spans S with the MST on S'. It is immediate that after the replacement, the new solution has the same connectivity of portals and terminal points as before. We define F' as the solution after doing this replacement for all simple squares.

F' is still a feasible solution. Furthermore, for every simple square R, if F is compatible with a subproblem (R, A, f, Π), then so does F'. By the construction of F', the definition of Steiner ratio α_2, and Theorem 2.2, we know that

$$w(F') \le \alpha_2 \cdot w(F) \le (1 + O(\varepsilon)) \cdot \alpha_2 \cdot \mathrm{OPT}, \tag{2}$$

where the last inequality holds as we condition on that the error bound in Theorem 2.2 holds.

Now we relate the algorithm's cost to $w(F')$. Fix a simple square R, and suppose (R, A, f, Π) is the subproblem that is compatible with $F'|_R$. Then, the components in $F'|_R$ can be described by a partition of the cells plus their connectivity to active portals. Such a subproblem, together with the partition, must be examined by the algorithm (in Line 17 to Line 29), and the MST value for each part is estimated in Line 24. Since the algorithm runs a DP using the estimated values, the final DP value is no worse than the DP value that is only evaluated from the subproblems that are compatible to F'. Recall that our estimation for each subproblem not only has a multiplicative error of $(1+\varepsilon)$ but also an additive error by Lemma 3.7. Therefore, by the fact that F' always uses MST to connect points in components of basic subproblems, it suffices to bound the *total* additive error for the estimation of the MST cost of the components of F'.

Fix a connected (global) component Q of F', and let $\mathcal{C}_Q \subseteq \mathcal{C}$ be the subset of colors that belongs to Q. By Lemma 3.7, for every basic subproblem (R, f, A, Π) that is compatible with F', and every component P of $Q|_R$, the additive error is at most $O\left(\frac{\mathrm{poly}(\varepsilon)}{k \log \Delta}\right) \cdot \mathrm{MST}(S)$, where S is the union of color classes that intersect P plus the active portals A. Observe that $\mathcal{C}_S \subseteq \mathcal{C}_Q$ (where $\mathcal{C}_S$ is the set of colors used in S), so S is a subset of the point set of Q (note that Q contains all portals in A as F' is a portal-respecting solution and the subproblem is compatible with F') and thus $\mathrm{MST}(S) \le \mathrm{MST}(Q)$, which implies

$$O\left(\frac{\mathrm{poly}(\varepsilon)}{k \log \Delta}\right) \cdot \mathrm{MST}(S) \le O\left(\frac{\mathrm{poly}(\varepsilon)}{k \log \Delta}\right) \cdot \mathrm{MST}(Q) \le O\left(\frac{\mathrm{poly}(\varepsilon)}{k \log \Delta}\right) \cdot w(Q).$$

Observe that for each simple square R, $Q|_R$ has at most $O(\varepsilon^{-2})$ local components, hence, summing over all local components of $Q|_R$ and all simple squares R, the total additive error is bounded by

$$\mathrm{poly}(\varepsilon^{-1}) \cdot O(k \log \Delta) \cdot O\left(\frac{\mathrm{poly}(\varepsilon)}{k \log \Delta}\right) \cdot w(Q) \le \varepsilon \cdot w(Q)\,,$$

where use that there are at most $O(k \log \Delta)$ simple squares by Lemma 3.3. Finally, summing over all components Q of F', we conclude that the total additive error is $\varepsilon \cdot w(F')$. Combining with Equation (2), we conclude the error guarantee. This finishes the proof of Theorem 1.1.

3.4 Proof of Lemma 3.7

We first restate the lemma for convenience.

Lemma 3.7. *There is an algorithm that for every $0 < \varepsilon < 1$, integer $k, \Delta \geq 1$, and every set of points $S \subseteq [\Delta]^2$ presented as a dynamic geometric stream, maintains a linear sketch of size $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$. For every query $(R, \{R_j\}_{j=1}^t, A)$ (provided after the stream ends) satisfying*

1. *R is a simple square, A is a subset of portals of R, and*
2. *$\{R_j\}_{j=1}^t \subseteq \mathsf{cell}(R)$,*

the algorithm computes from the linear sketch a real number E such that with probability at least $1 - \exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta))$,

$$\mathrm{MST}(P) \leq E \leq (1+\varepsilon) \cdot \mathrm{MST}(P) + O\left(\frac{\mathrm{poly}(\varepsilon)}{k \log \Delta}\right) \cdot \mathrm{MST}(S)\,,$$

where $P = \left(\bigcup_{j=1}^t R_j\right) \cup A$. The algorithm runs in time $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$ per update and the query time is also $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$.

The general proof strategy is similar to that of [FIS08], hence we start with a review of [FIS08] (with some adjustments suitable for our setting), and our description is with respect to a generic point set $V \subseteq [\Delta]^2$.

Review of [FIS08]. A key observation is that the cost of the MST of a point set $V \subseteq [\Delta]^2$ can be related to the number of components in the metric threshold graphs of different scales; a similar observation was first given in [CRT05]. In particular, let G_V be the complete metric graph on V, i.e., the vertex set is V, the edge set is $\{\{u, v\} : u \neq v \in V\}$, and the edge weights are given by $\mathrm{dist}(\cdot, \cdot)$. For $i \geq 0$, let $G_V^{(i)}$ be the $(1+\varepsilon)^i$-threshold graph, which only consists of edges of G_V that have weight at most $(1+\varepsilon)^i$. Let $c_V^{(i)}$ be the number of connected components in $G_V^{(i)}$. Then for every $W \geq \mathrm{diam}(V)$ that is a power of $(1+\varepsilon)$,

$$\mathrm{MST}(V) \leq c_V^{(0)} - W + \varepsilon \cdot \sum_{i=0}^{\log_{1+\varepsilon} W - 1} (1+\varepsilon)^i \cdot c_V^{(i)} \leq (1+\varepsilon) \cdot \mathrm{MST}(V). \tag{3}$$

It remains to estimate $c_V^{(i)}$ for $0 \leq i \leq \log_{1+\varepsilon} W - 1$. For each i, Frahling et al. [FIS08] consider the subdivision of $[\Delta]^2$ into squares of side-length $(1+\varepsilon)^{i'}$ for a small enough i' and round the input points to to grid points formed by centers of these squares. Namely, i' is the largest integer satisfying $(1+\varepsilon)^{i'} \leq O(\varepsilon \cdot (1+\varepsilon)^i)$. However, this is not convenient in our setting as we need to restrict the MST query to a subset of quad-tree squares (with side-lengths of powers of 2). As it is not possible to "align" squares of size $(1+\varepsilon)^{i'}$ with quad-tree squares, the aforementioned rounding could move a point not relevant to the MST query to a quad-tree square of the query (thus making it appear relevant), or vice versa.

To avoid this issue, we adjust the rounding for MST cost estimation so that the grid points are centers of squares in the randomly-shifted quad-tree we use in the DP computation. Recall from Section 2 that the $\mathcal{G}_{i'}$ is the set of centers of all level-i' quad-tree squares. Namely, to estimate $c_V^{(i)}$, we round input points to the $2^{i'}$-grid $\mathcal{G}_{i'}$, where i' is the largest integer such that $2^{i'} \leq O(\varepsilon \cdot (1+\varepsilon)^i)$, that is,

$$i' := \log_2 O(\varepsilon \cdot (1+\varepsilon)^i)\,. \tag{4}$$

We require the constant hidden in the O notation to be sufficiently small so that an inequality in (8) holds. (If $i' < 0$ according to this definition, then we can of course take $i' = 0$ and no rounding is needed as we reach the granularity of the input data.) Our rounding is only finer compared to [FIS08] (i.e., to centers of smaller squares), but within a constant factor, so the space bound remains asymptotically the same. Note that while the squares of $\mathcal{G}_{i'}$ are from the quad-tree, we still need to show that we do not round to centers of larger squares than the cells of the MST query, which will imply that indeed, our rounding does not move a point irrelevant for the query to a cell of the query, or vice versa. We remark that for several consecutive indexes $i \in [0, \log_{1+\varepsilon} W - 1]$, the index i' could be of the same value.

Similarly as in [FIS08], for each i, define a "rounded" metric graph $\mathring{G}_V^{(i)}$ as follows.

1. Move each point $x \in V$ to the center $y \in \mathcal{G}_{i'}$ of the quad-tree square that x belongs to, where i' is defined as in (4).
2. The vertex set of $\mathring{G}_V^{(i)}$ consists of *non-empty* grid points in $\mathcal{G}_{i'}$, i.e., for each vertex v at least one point in V was moved/rounded to the grid point corresponding to v, and the edge set consists of pairs of vertices that are of distance at most $(1+\varepsilon)^i$.

We refer to the vertices of $\mathring{G}_V^{(i)}$ as *non-empty* grid points. Let $\mathring{c}_V^{(i)}$ be the number of components in $\mathring{G}_V^{(i)}$, and let $\mathring{n}^{(i)}$ be the number of vertices in $\mathring{G}_V^{(i)}$. As shown in [FIS08], $c_V^{(i)}$ is well approximated by $\mathring{c}_V^{(i)}$, which we restate in Lemma 3.8. Strictly speaking, Lemma 3.8 ([FIS08, Claim 4.1]) is for the rounding to centers of squares with side-length $O(\varepsilon \cdot (1+\varepsilon)^i)$, but its proof more generally applies to any rounding which moves points by at most $O(\varepsilon \cdot (1+\varepsilon)^i)$ and thus, to our rounding as well.

Lemma 3.8 ([FIS08, Claim 4.1])**.** *For every i, $c_V^{(i+1)} \le \mathring{c}_V^{(i)} \le c_V^{(i-2)}$.*

Therefore, we focus on estimating $\mathring{c}_V^{(i)}$'s. As in [FIS08], it suffices to account for *small* components that have at most $O(\varepsilon^{-2} \log \Delta)$ non-empty grid points for each i. The reason is that the number of large components that have more than $\varepsilon^{-2} \log \Delta$ points is at most $O(\varepsilon^2 / \log \Delta \cdot \mathring{n}_V^{(i)})$, which contributes $O(\varepsilon^2 / \log \Delta \cdot (1+\varepsilon)^i \cdot \mathring{n}_V^{(i)})$ in Equation (3). This contribution can be bounded using the following lemma[3].

Lemma 3.9 (Lower bound on MST [FIS08, Lemma 4])**.** *For every i, $\mathrm{MST}(V) \ge \Omega((1+\varepsilon)^i \cdot \mathring{n}_V^{(i)})$.*

To estimate the number of small components, we use the BFS algorithm with a stochastic stopping condition; this idea was first applied in [CRT05]. In particular, for each i, $\mathrm{poly}(\varepsilon^{-1} \log \Delta)$ samples are taken from the point set of $\mathring{G}_V^{(i)}$, which may be efficiently maintained and sampled using ℓ_0-samplers. After that, we perform a stochastic-stopping BFS in $\mathring{G}_V^{(i)}$ starting from each sampled point and using a random number of steps (but at most $O(\varepsilon^{-2} \log \Delta)$ steps). An estimate for the number of small components, and thus for $\mathring{c}_V^{(i)}$, is computed using the outcome of each BFS, i.e., whether the whole component is discovered or not. The random exploration is made possible by the following modified ℓ_0-sampler designed in [FIS08], which also returns the non-empty neighborhood when sampling a point. See also a survey about ℓ_0-samplers by Cormode and Firmani [CF14].

Lemma 3.10 (ℓ_0-Sampler with neighborhood information [FIS08, Corollary 3])**.** *There is an algorithm that for $\delta > 0$, integer $\rho, \Delta \ge 1$, every set of points $S \subseteq [\Delta]^2$ presented as a dynamic*

[3] Note that Lemma 4 in [FIS08] proves the lower bound only for the case when $W \ge \Omega(\varepsilon \cdot (1+\varepsilon)^i)$, however, using the argument of the first case of their proof implies our bound.

geometric stream, succeeds with probability at least $1-\delta$ and, conditioned on it succeeding, returns a point $p \in S$ such that for every $s \in S$ it holds that $\Pr[p = s] = 1/|S|$. Moreover, if the algorithm succeeds, it also returns all points from $s \in S$ such that $\mathrm{dist}(p, s) \le \rho$. The algorithm has space and both update and query times bounded by $\mathrm{poly}(\rho \cdot \varepsilon^{-1} \cdot \log \Delta \cdot \log \delta^{-1})$, and its memory contents is a linear sketch of S.

Handling subset queries. In our case, we need to apply Equation (3) with $V = P$, recalling that P is the point set of the query. To pick W in (3), we need an upper bound on $\mathrm{diam}(P)$. Suppose in the query, the simple square R is of level-i_0, then $\mathrm{diam}(P) \le 2^{i_0}$. Hence, we pick W to be the smallest power of $(1+\varepsilon)$ that is no smaller than 2^{i_0}, which implies

$$2^{i_0} \le W \le (1+\varepsilon) \cdot 2^{i_0}\,. \tag{5}$$

However, the query set P is given after the stream, and it is provided in a compact form as a union of several cells and portals. On the other hand, what we can maintain is only a sketch for S (the whole point set). Therefore, the above idea from [FIS08] cannot immediately solve our problem, and as discussed in Section 3.2.3, we cannot count on adapting ℓ_0-samplers to directly work on a subset unknown in advance either. Hence, we still have to maintain the sketch for the whole S, and query the ℓ_0-samplers on S. Then naturally, the only way out seems to filter out the irrelevant sampled non-empty grid points, which are those not corresponding to the query set, and then we simulate the MST algorithm on the local query instance, by using those relevant samples. Apart from filtering out irrelevant samples, we also need to restrict the neighborhoods (from Lemma 3.10) to the query cells, i.e., filter out irrelevant point from the neighborhoods.

At a first glance, this idea makes sense and should generally work. However, we now discuss an outstanding issue and present our new technical ideas for resolving it.

Additive error and failure probability. As mentioned in Section 3.2.3, the sampling idea may not work if P has relatively few relevant grid points $\mathcal{G}_{i'}$, since it is difficult to collect enough relevant grid points using a small number of samples, and in this case we need to suffer an additive error. Luckily, we show that if the number of relevant grid points is indeed small relative to that of S, then the contribution of these points to the global MST cost is small after all; see Lemma 3.11. This yields a small additive error that can be eventually charged to the global MST of S. A technical issue here is that the algorithm does not know whether or not P contains very few relevant grid points in advance. Hence, we introduce an additional sampling step to estimate this density, and we use this estimate ($\mathcal{S}^{(i)}$ in Algorithm 6) to guide the algorithm. This works well for the algorithm, but in the analysis, the random estimate cannot assert for sure whether or not P contains relatively few relevant grid points. We thus need to do the case analysis based on the actual number of grid points in P, which is not directly aligned with the case separation of the algorithm.

Finally, since we need to apply the query on all subproblems (in Section 3.3), this requires that both the success probability and the additive error are as small as $\frac{1}{k}$ (ignoring other factors). In the original analysis by [FIS08], Chebyshev's inequality was used to bound the failure probability, which in our case only suffices for a k^4 space bound. Hence, we instead apply Hoeffding's inequality, and this saves a factor of k in space compared to the original calculation in [FIS08].

Algorithm description. We state the complete algorithm for maintaining the linear sketch in Algorithm 5, and the query algorithm in Algorithm 6. In a nutshell, maintaining the sketch in Algorithm 5 merely requires updating extended ℓ_0-samplers of Lemma 3.10 (including the associated neighborhood information) together with keeping track of the ℓ_0 norm for each level $0, \ldots, \log_2 \Delta$

(more precisely, for each threshold $i = 1, \ldots, \log_{1+\varepsilon} \Delta$, we maintain the ℓ_0-samplers and the ℓ_0 norm in level i' of the quad-tree, where i' is defined as in (4)). Note that Algorithm 5 indeed outputs a linear sketch, since the extended ℓ_0-samplers of Lemma 3.10 is a linear sketch and the ℓ_0-norm estimators as well (cf. [KNW10]).

In the query procedure (Algorithm 6), for each threshold $i = 1, \ldots, \log_{1+\varepsilon} \Delta$, we start by checking whether or not the query contains relatively few non-empty grid points, for which we use the first σ of ℓ_0-samplers; namely, we check if at least κ of sampled grid points are relevant for the query, and if not, our estimate for the number of components on that level is simply 0. Otherwise, the query contains relatively large number of non-empty grid points with high probability, and we query the remaining σ of ℓ_0-samplers, execute the stochastic-stopping BFS from each relevant sampled grid point, and use the outcomes of the BFS (i.e. whether or not the whole component was discovered) to estimate the number of components on that level, as described above.

Algorithm 5 Algorithm for maintaining sketches for Lemma 3.7

```
procedure MST-SKETCH($S$)    ▷ $S$ is provided as a dynamic geometric stream
    let $\kappa \leftarrow k \log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta)$, $\sigma \leftarrow k\,\mathrm{poly}(\varepsilon^{-1} \log \Delta) \cdot \kappa$, $\Gamma \leftarrow \varepsilon^{-2} \log \Delta$    ▷ as in (6)
    for $i \leftarrow 0, \ldots, \log_{1+\varepsilon} \Delta$ do
        let $i'$ be defined as in (4), i.e., the largest integer such that $2^{i'} \leq O(\varepsilon \cdot (1+\varepsilon)^i)$
        initialize $2\sigma$ extended $\ell_0$-samplers $\{\mathcal{K}_j^{(i)}\}_j$ of Lemma 3.10 with frequency vector indexed
    by $\mathcal{G}_{i'}$ and with neighborhoods containing all grid points from $\mathcal{G}_{i'}$ at distance at most $\Gamma \cdot (1+\varepsilon)^i$
        initialize an $\ell_0$-norm estimator $\mathcal{N}^{(i)}$ (cf. [KNW10]) for the number of non-empty grid
    points in $\mathcal{G}_{i'}$ with error parameter $\varepsilon$
    end for
    for each insertion/deletion of point $x$ do
        for $i \leftarrow 0, \ldots, \log_{1+\varepsilon} \Delta$ do
            let $i'$ be defined as in (4)
            let $y \in \mathcal{G}_{i'}$ be the center of the level-$i'$ quad-tree square containing $x$
            increase/decrease the frequency of $y$ by 1 for estimator $\mathcal{N}^{(i)}$
            for all $\ell_0$-samplers $\{\mathcal{K}_j^{(i)}\}_j$, increase/decrease by 1 the following:
              • the frequency of $y$, and
              • the frequency of $y$ in the neighborhood of any grid point $z \in \mathcal{G}_{i'}$ with $\mathrm{dist}(z, y) \leq \Gamma \cdot (1+\varepsilon)^i$
                (cf. [FIS08])
        end for
    end for    ▷ stream of $S$ terminates
end procedure
```

Both Algorithms 5 and 6 use the following parameters

$$\kappa = k \log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta), \quad \sigma = k\,\mathrm{poly}(\varepsilon^{-1} \log \Delta) \cdot \kappa, \quad \Gamma = \varepsilon^{-2} \cdot \log \Delta. \tag{6}$$

We additionally require that

$$\frac{\kappa^2}{\sigma \cdot \Gamma^2} \geq \log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta)\,, \tag{7}$$

where $\mathrm{poly}(\varepsilon^{-1} \log \Delta)$ in the RHS is the same as in the failure probability of Lemma 3.7. Note that $\sigma = k^2 \log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta)$.

Algorithm 6 Algorithm for answering queries of Lemma 3.7

```
procedure QUERY($R, \{R_1, \dots, R_t\}, A$)    ▷ assume the access to sketches in Algorithm 5
    let $\kappa, \sigma, \Gamma$ be the same parameters as in Algorithm 5
    let $W$ be as in 5, i.e., the smallest power of $(1+\varepsilon)$ that is no smaller than $2^{i_0}$, where $i_0$ is the level of $R$ in the quad-tree
    for $i \leftarrow 0, \dots, \log_{1+\varepsilon} W$ do
        let $i'$ be as in (4), i.e., the largest integer such that $2^{i'} \leq O(\varepsilon \cdot (1+\varepsilon)^i)$
        query $\mathcal{N}^{(i)}$ and record the value as an estimate $\widetilde{n}_S^{(i)}$
        query the first $\sigma$ of $\ell_0$-samplers $\{\mathcal{K}_j^{(i)}\}_j$, and suppose the set of uniformly sampled points is $\{p_j\}_{j=1}^{\sigma}$ and $\{U_j)\}_{j=1}^{\sigma}$ are the associated neighborhoods in grid $\mathcal{G}_{i'}$
        let $\mathcal{S}^{(i)} \leftarrow \{p_j : \text{IS-RELEVANT}(i',\ p_j\ ) = \text{TRUE}\}$ be the subset relevant to the query
        if $|\mathcal{S}^{(i)}| < \kappa$ then
            define estimator $\widetilde{c}_P^{(i)} \leftarrow 0$
        else
            query the remaining $\sigma$ of $\ell_0$-samplers $\{\mathcal{K}_j^{(i)}\}_j$, and use the same notations $\{p_j\}_{j=1}^{\sigma}$ as well as $\{(U_j)\}_{j=1}^{\sigma}$, to denote the outcome
            for each $p_j$, let $\beta_{p_j} \leftarrow \text{BFS}(i, i', p_j, U_j)$
            define estimator $\widetilde{c}_P^{(i)} \leftarrow (\widetilde{n}_S^{(i)}/\sigma) \cdot \sum_{j=1}^{\sigma} I(\text{IS-RELEVANT}(i', p_j) = \text{TRUE}) \cdot \beta_{p_j}$
        end if
    end for
    query the original MST sketch algorithm of [FIS08], and let $\widetilde{\mathrm{MST}}(S)$ be the outcome
    return $\widetilde{\mathrm{MST}} \leftarrow \Theta\left(\frac{\mathrm{poly}(\varepsilon)}{k \log \Delta}\right) \widetilde{\mathrm{MST}}(S) + \widetilde{c}_P^{(0)} - W + \varepsilon \cdot \sum_{i=0}^{\log_{1+\varepsilon} W} (1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)}$
            ▷ the $\widetilde{\mathrm{MST}}(S)$ term is used to make sure $\widetilde{\mathrm{MST}}$ is never an underestimation w.h.p.
end procedure
procedure BFS($i, i', p, U$)
    let $U' \leftarrow \{q \in U : \text{IS-RELEVANT}(i',\ q\ ) = \text{TRUE}\}$ be the restriction of the neighborhood of $p$ to the query    ▷ recall that $U$ contains all non-empty grid points of $\mathcal{G}_{i'}$ at distance $\leq \Gamma \cdot (1+\varepsilon)^i$ from $p$
    sample integer $Y$ according to distribution $\Pr[Y \geq m] = \frac{1}{m}$
    if $Y \geq \Gamma$ or the component in the $(1+\varepsilon)^i$-threshold graph on $U'$ that contains $p$ has more than $Y$ vertices, set $\beta \leftarrow 0$, otherwise set $\beta \leftarrow 1$
    return $\beta$
end procedure
procedure IS-RELEVANT($i', p$)
    return TRUE if (i) $\exists x \in A$ s.t. portal $x$ belongs to the level-$i'$ quad-tree square corresponding to $p$, or (ii) $\exists R_j$ (which is a query cell) s.t. $p \in R_j$; otherwise return FALSE
end procedure
```

Space and time analysis. As can be seen from Algorithm 5, the space is dominates by the 2σ extended ℓ_0-samplers of Lemma 3.10 with $\Gamma = \varepsilon^{-2} \log \Delta$ for each $0 \leq i \leq \log_{1+\varepsilon} W$. Thus, the space bound follows from Lemma 3.10 and from the value of σ, defined in (6).

The update time is also dominated by maintaining 2σ extended ℓ_0-samplers of Lemma 3.10, which can be done in time $\mathrm{poly}(\log k \cdot \Gamma \cdot \varepsilon^{-1} \log \Delta))$ for each sampler, including the updates to the associated neighborhoods. Thus, the total update time is $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$. Finally, the query time is bounded by querying all ℓ_0-samplers and executing the stochastic-stopping BFS from at most σ sampled points, each with at most Γ steps, which overall takes time of $k^2 \cdot \mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$.

3.4.1 Error Analysis

We analyze the error of the query procedure (Algorithm 6), using the notation defined in Algorithms 5 and 6. First, we show that procedure Is-Relevant captures the points in P exactly, i.e., for every i and $p \in \mathring{G}_S^{(i)}$, it holds that $p \in \mathring{G}_P^{(i)}$, i.e., it is relevant to the query P, if and only if Is-Relevant(i', p) returns TRUE. As we round to the centers of level-i' quad-tree squares, it suffices to shows that these level-i' squares are not larger than the cells of the query. Recall that the simple square containing these cells is of level i_0 and that $i \le \log_{1+\varepsilon} W \le \log_{1+\varepsilon}((1+\varepsilon) \cdot 2^{i_0})$ by (5). As the cells have side-length (at least) $\Theta(2^{i_0} \cdot \varepsilon)$ (see Section 2.2), we have

$$2^{i'} \le O(\varepsilon \cdot (1+\varepsilon)^i) \le O(\varepsilon \cdot (1+\varepsilon) \cdot 2^{i_0}) \le \Theta(2^{i_0} \cdot \varepsilon)\,, \tag{8}$$

using the definition of i' in (4) and that the constant hidden in the O notation in (4) is sufficiently small, compared to the constant hidden in Θ in cell side-length. This enables us to filter out the irrelevant samples from S and work only on $\mathring{G}_P^{(i)}$, i.e., the points in P rounded to the $2^{i'}$-grid $\mathcal{G}_{i'}$.

Fix some i. We bound the error for the estimators $\widetilde{c}_P^{(i)}$. First, by the guarantee of the ℓ_0-norm estimator $\mathcal{N}^{(i)}$ (see Line 6), we know that it uses space $\mathrm{poly}(\log k \cdot \varepsilon^{-1} \log \Delta)$ to achieve with probability at least $1 - \exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta))$,

$$\widetilde{n}_S^{(i)} \in (1 \pm \varepsilon) \cdot \mathring{n}_S^{(i)}\,. \tag{9}$$

We assume this happens, and the probability that it does not happen can be charged to the total failure probability. Similarly, we assume the success of all other sketches, namely $\mathcal{K}_j^{(i)}$'s, and we require the failure probability to be at most $\exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta))$.

We have two cases for $\widetilde{c}_P^{(i)}$ in Algorithm 6 depending on whether or not we collect enough samples $\mathcal{S}^{(i)}$ that pass the relevance test. However, whether or not enough samples are collected is a random event, which is not easy to handle if we do the case analysis on it. Therefore, we turn our attention to a tightly related quantity $\mathring{n}_P^{(i)} / \mathring{n}_S^{(i)}$, which is the fraction of non-empty grid points of $\mathcal{G}_{i'}$ in the query instance P. We analyze in Lemma 3.11 the error for the estimator when this fraction is small (and most likely, not enough samples are collected), and then, in Lemma 3.12, the estimation when $\mathring{n}_P^{(i)} / \mathring{n}_S^{(i)}$ is large.

Lemma 3.11. *For every $\lambda > 0$, if $\mathring{n}_P^{(i)} \le \lambda \cdot \mathring{n}_S^{(i)}$, then $(1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} \le O(\lambda \cdot \mathrm{MST}(S))$, which implies that the estimator $\widetilde{c}_P^{(i)} = 0$ satisfies*

$$(1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} - O(\lambda \cdot \mathrm{MST}(S)) \le (1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)} \le (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)}.$$

Proof. Using $\mathring{c}_P^{(i)} \le \mathring{n}_P^{(i)}$ and the condition of the lemma, we get

$$(1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} \le (1+\varepsilon)^i \cdot \mathring{n}_P^{(i)} \le (1+\varepsilon)^i \cdot \lambda \cdot \mathring{n}_S^{(i)} \le O(\lambda \cdot \mathrm{MST}(S))\,,$$

where the last inequality follows from Lemma 3.9. □

Lemma 3.12. *For every $\lambda > 0$, if $\mathring{n}_P^{(i)} \ge \lambda \cdot \mathring{n}_S^{(i)}$, then with probability at least $1 - \exp\left(-\Omega(\sigma) \cdot \left(\frac{\lambda}{\Gamma}\right)^2\right)$, the estimator $\widetilde{c}_P^{(i)}$ in line Line 14 of Algorithm 6 satisfies*

$$|\widetilde{c}_P^{(i)} - \mathring{c}_P^{(i)}| \le O(\varepsilon) \cdot \mathring{c}_P^{(i)} + O\left(\frac{\mathrm{MST}(P)}{\Gamma \cdot (1+\varepsilon)^i}\right).$$

Proof. We use some of the notations from Algorithm 6, and suppose p_j and U_j are those from Line 12. For the ease of notation, let $\mathcal{E}_j$ be the event that $\textsc{Is-Relevant}(i', p_j) = \text{TRUE}$. Using a similar calculation as in [FIS08], we observe that for every p_j,

$$\begin{aligned}
\mathbb{E}[I(\mathcal{E}_j) \cdot \beta_{p_j}] &= \Pr[\mathcal{E}_j] \cdot \Pr[\beta_{p_j} = 1 \mid \mathcal{E}_j] \\
&= \frac{\mathring{n}_P^{(i)}}{\mathring{n}_S^{(i)}} \cdot \sum_{\text{component H in } \mathring{G}_P^{(i)}} \Pr[p_j \in H \mid \mathcal{E}_j] \cdot \Pr[Y \geq |H| \wedge Y < \Gamma] \\
&= \frac{\mathring{n}_P^{(i)}}{\mathring{n}_S^{(i)}} \cdot \sum_{\text{component H in } \mathring{G}_P^{(i)}} \frac{|H|}{\mathring{n}_P^{(i)}} \cdot \Pr[Y \geq |H| \wedge Y < \Gamma] \\
&= \frac{1}{\mathring{n}_S^{(i)}} \cdot \sum_{\text{component H in } \mathring{G}_P^{(i)}} |H| \cdot \Pr[Y \geq |H| \wedge Y < \Gamma]. \qquad (10)
\end{aligned}$$

Then, we get an upper bound on $\mathbb{E}[\widetilde{c}_P^{(i)}]$ using (9),

$$\begin{aligned}
\mathbb{E}[\widetilde{c}_P^{(i)}] &= \frac{\widetilde{n}_S^{(i)}}{\sigma} \sum_{j=1}^{\sigma} \mathbb{E}[I(\mathcal{E}_j) \cdot \beta_{p_j}] \\
&= \frac{\widetilde{n}_S^{(i)}}{\sigma \mathring{n}_S^{(i)}} \sum_{j=1}^{\sigma} \sum_{\text{component } H \text{ in } \mathring{G}_P^{(i)}} |H| \cdot \Pr[Y \geq |H| \wedge Y < \Gamma] \\
&\leq \frac{1+\varepsilon}{\sigma} \sum_{j=1}^{\sigma} \sum_{\text{component } H \text{ in } \mathring{G}_P^{(i)}} |H| \cdot \Pr[Y \geq |H|] \\
&= (1+\varepsilon) \cdot \mathring{c}_P^{(i)},
\end{aligned}$$

where the last step follows from the distribution of Y. Similarly, we also get a lower bound

$$\begin{aligned}
\mathbb{E}[\widetilde{c}_P^{(i)}] &= \frac{\widetilde{n}_S^{(i)}}{\sigma} \sum_{j=1}^{\sigma} \mathbb{E}[I(\mathcal{E}_j) \cdot \beta_{p_j}] \\
&= \frac{\widetilde{n}_S^{(i)}}{\sigma \mathring{n}_S^{(i)}} \sum_{j=1}^{\sigma} \sum_{\text{component } H \text{ in } \mathring{G}_P^{(i)}} |H| \cdot \Pr[Y \geq |H| \wedge Y < \Gamma] \\
&\geq \frac{1-\varepsilon}{\sigma} \sum_{j=1}^{\sigma} \sum_{\text{component } H \text{ in } \mathring{G}_P^{(i)}} |H| \cdot \left(\frac{1}{|H|} - \frac{1}{\Gamma} \right) \\
&\geq (1-\varepsilon) \cdot \mathring{c}_P^{(i)} - \frac{1-\varepsilon}{\Gamma} \cdot \mathring{n}_P^{(i)}.
\end{aligned}$$

Next, we apply Hoeffding's inequality. To this end, consider random variables $Z_j = \widetilde{n}_S^{(i)}/\sigma \cdot I(\mathcal{E}_j) \cdot \beta_{p_j}$ for $j = 1, \ldots, \sigma$ and note that they are independent and we have that $\widetilde{c}_P^{(i)} = \sum_{j=1}^{\sigma} Z_j$. Therefore, by Hoeffding's inequality, it holds that

$$\begin{aligned}\Pr\left[|\widetilde{c}_P^{(i)} - \mathbb{E}[\widetilde{c}_P^{(i)}]| \geq \Omega\left(\frac{\mathring{n}_P^{(i)}}{\Gamma}\right)\right] &\leq 2\exp\left(-2\cdot\Omega\left(\frac{\mathring{n}_P^{(i)}}{\Gamma}\right)^2\cdot\frac{1}{\sigma}\cdot\left(\frac{\sigma}{\widetilde{n}_S^{(i)}}\right)^2\right)\\ &= \exp\left(-\Omega(\sigma)\cdot\left(\frac{\mathring{n}_P^{(i)}}{\Gamma\widetilde{n}_S^{(i)}}\right)^2\right) \leq \exp\left(-\Omega(\sigma)\cdot\left(\frac{\lambda}{\Gamma}\right)^2\right).\end{aligned}$$

where in the second inequality, we use $\mathring{n}_P^{(i)} \geq \lambda\cdot\mathring{n}_S^{(i)} \geq \lambda\cdot(1-\varepsilon)\cdot\widetilde{n}_S^{(i)}$ by the assumption of the lemma and by (9). Combining the expectation bound and the above concentration inequality, we conclude that with probability at least $1 - 2\exp\left(-\Omega(\sigma)\cdot\left(\frac{\lambda}{\Gamma}\right)^2\right)$,

$$|\widetilde{c}_P^{(i)} - \mathring{c}_P^{(i)}| \leq O(\varepsilon)\cdot\mathring{c}_P^{(i)} + O\left(\frac{\mathring{n}_P^{(i)}}{\Gamma}\right) \leq O(\varepsilon)\cdot\mathring{c}_P^{(i)} + O\left(\frac{\mathrm{MST}(P)}{\Gamma\cdot(1+\varepsilon)^i}\right),$$

where the last inequality follows from Lemma 3.9. □

Next, we do the following case analysis. Let $\lambda_1 := \frac{\kappa}{2\sigma}$, and $\lambda_2 := \frac{2\kappa}{\sigma}$; note that $\lambda_1 \leq \lambda_2$.

Case I: $\mathring{n}_P^{(i)} < \lambda_1\cdot\mathring{n}_S^{(i)}$. We claim that with high probability, $|\mathcal{S}^{(i)}| < \kappa$. This can be done by using Hoeffding's inequality, as shown in Claim 3.13.

Claim 3.13. *If $\mathring{n}_P^{(i)} < \lambda_1\cdot\mathring{n}_S^{(i)}$, then* $\Pr[|\mathcal{S}^{(i)}| \geq \kappa] \leq \exp(-\log k\cdot\mathrm{poly}(\varepsilon^{-1}\log\Delta))$.

Proof. Let Z_j be the $\{0,1\}$ random variable, that takes 1 if Is-Relevant(i', p_j) = TRUE. Then

$$\Pr[Z_j = 1] = \frac{\mathring{n}_P^{(i)}}{\mathring{n}_S^{(i)}}.$$

So $|\mathcal{S}^{(i)}| = \sum_{j=1}^{\sigma} Z_j$, and hence,

$$\mathbb{E}[\mathcal{S}^{(i)}] = \frac{\sigma\mathring{n}_P^{(i)}}{\mathring{n}_S^{(i)}} < \sigma\lambda_1 = \frac{\kappa}{2}.$$

By Hoeffding's inequality,

$$\begin{aligned}\Pr[|\mathcal{S}^{(i)}| \geq \kappa] &= \Pr[|\mathcal{S}^{(i)}| - \mathbb{E}[|\mathcal{S}^{(i)}|] \geq \kappa - \mathbb{E}[\mathcal{S}^{(i)}]]\\ &\leq \Pr[|\mathcal{S}^{(i)}| - \mathbb{E}[|\mathcal{S}^{(i)}|] \geq \frac{\kappa}{2}]\\ &\leq \exp(-\frac{\kappa^2}{2\sigma}) = \exp(-\log k\cdot\mathrm{poly}(\varepsilon^{-1}\log\Delta)),\end{aligned}$$

where the last inequality follows from (7). □

Then, we assume $|\mathcal{S}^{(i)}| < \kappa$ happens in Case I, and by using Lemma 3.11 with $\lambda = \lambda_1$, we conclude that the estimator $\widetilde{c}_P^{(i)} = 0$ satisfies

$$(1+\varepsilon)^i\cdot\mathring{c}_P^{(i)} - O(\lambda_1\,\mathrm{MST}(S)) \leq (1+\varepsilon)^i\cdot\widetilde{c}_P^{(i)} \leq (1+\varepsilon)^i\cdot\mathring{c}_P^{(i)}. \qquad (11)$$

The case $|\mathcal{S}^{(i)}| \geq \kappa$ only happens with a very small probability, and we charge it to the total failure probability of Lemma 3.7.

Case II: $\mathring{n}_P^{(i)} > \lambda_2 \cdot \mathring{n}_S^{(i)}$. Similar to Case I, we claim that with probability $1 - \exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta))$, $|\mathcal{S}^{(i)}| \geq \kappa$. This can be done again by using Hoeffding's inequality, and we omit the details since it is very similar to that in Claim 3.13. Then, assuming $|\mathcal{S}^{(i)}| \geq \kappa$ and using Lemma 3.12 with $\lambda = \lambda_2$, we conclude that with probability $1 - \exp\left(-\Omega(\sigma) \cdot \left(\frac{\lambda_2}{\Gamma}\right)^2\right)$, the estimator in line Line 14 of Algorithm 6 satisfies

$$|\widetilde{c}_P^{(i)} - \mathring{c}_P^{(i)}| \leq O(\varepsilon) \cdot \mathring{c}_P^{(i)} + O\left(\frac{\mathrm{MST}(P)}{\Gamma \cdot (1+\varepsilon)^i}\right). \tag{12}$$

Case III: None of the other two cases happens, so $\lambda_1 \cdot \mathring{n}_S^{(i)} \leq \mathring{n}_P^{(i)} \leq \lambda_2 \cdot \mathring{n}_S^{(i)}$. Then we cannot decide with high probability which type of estimate for $\widetilde{c}_P^{(i)}$ the algorithm uses. However, since we have both an upper and lower bound for $\mathring{n}_P^{(i)} / \mathring{n}_S^{(i)}$, Lemmas 3.11 and 3.12 can both be applied with a reasonable guarantee. In particular, we apply Lemma 3.11 with $\lambda = \lambda_2$, which for the estimator $\widetilde{c}_P^{(i)} = 0$ implies

$$(1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} - O(\lambda_2 \, \mathrm{MST}(S)) \leq (1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)} \leq (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)}.$$

Next, Lemma 3.12 with $\lambda = \lambda_1$ yields with probability at least $1 - \exp\left(-\Omega(\sigma) \cdot \left(\frac{\lambda_1}{\Gamma}\right)^2\right)$,

$$|\widetilde{c}_P^{(i)} - \mathring{c}_P^{(i)}| \leq O(\varepsilon) \cdot \mathring{c}_P^{(i)} + O\left(\frac{\mathrm{MST}(P)}{\Gamma \cdot (1+\varepsilon)^i}\right)$$

for the estimator in line Line 14 of Algorithm 6. Since we do not know which estimator the algorithm actually uses, to bound the error, we need to take the worse bound for both directions of these two inequalities, and for the failure probability as well. There, it always holds with probability at least $1 - \exp\left(-\Omega(\sigma) \cdot \left(\frac{\lambda_1}{\Gamma}\right)^2\right)$,

$$(1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)} \leq (1 + O(\varepsilon)) \cdot (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} + O\left(\frac{\mathrm{MST}(P)}{\Gamma}\right), \tag{13}$$

and that

$$(1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)} \geq (1 - O(\varepsilon)) \cdot (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} - O\left(\frac{\mathrm{MST}(P)}{\Gamma}\right) - O(\lambda_2 \, \mathrm{MST}(S)). \tag{14}$$

Conclusion of the error analysis. Overall, we bound $\widetilde{c}_P^{(i)}$ using the worse bound in both directions (similarly as in Case III). Namely, using Equations (11) to (14) and the union bound, with failure probability at most

$$\begin{aligned} & O(\log_{1+\varepsilon} W) \cdot \left(\exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta)) + \exp\left(-\Omega(\sigma) \cdot \left(\frac{\lambda_1}{\Gamma}\right)^2\right)\right) \\ & \leq \exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta)) + \exp\left(-\Omega\left(\frac{\kappa^2}{\sigma \Gamma^2}\right)\right) \\ & \leq \exp(-\log k \cdot \mathrm{poly}(\varepsilon^{-1} \log \Delta)), \end{aligned} \tag{15}$$

where the inequality follows from (7), for any $i = 0, \ldots, \log_{1+\varepsilon} W$, it holds that

$$(1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)} \le (1+O(\varepsilon)) \cdot (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} + O\left(\frac{\mathrm{MST}(P)}{\Gamma}\right), \tag{16}$$

and that

$$(1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)} \ge (1-O(\varepsilon)) \cdot (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} - O\left(\frac{\mathrm{MST}(P)}{\Gamma}\right) - O(\lambda_2 \, \mathrm{MST}(S)). \tag{17}$$

Summing over i, with failure probability bounded as in (15), we have

$$\begin{aligned}
&\sum_{i=0}^{\log_{1+\varepsilon} W} (1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)} \\
&\qquad \ge \sum_{i=0}^{\log_{1+\varepsilon} W} (1-O(\varepsilon)) \cdot (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} - O\left(\frac{\mathrm{MST}(P)}{\Gamma}\right) - O(\lambda_2 \, \mathrm{MST}(S)) \\
&\qquad \ge (1-O(\varepsilon)) \cdot \left(\sum_{i=1}^{\log_{1+\varepsilon} W} (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)}\right) - O\left(\frac{\log\Delta}{\varepsilon\Gamma}\right) \cdot \mathrm{MST}(P) - O\left(\frac{\lambda_2 \log\Delta}{\varepsilon}\right) \cdot \mathrm{MST}(S) \\
&\qquad \ge (1-O(\varepsilon)) \cdot \left(\sum_{i=1}^{\log_{1+\varepsilon} W} (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)}\right) - O(\varepsilon) \cdot \mathrm{MST}(P) - O\left(\frac{\kappa \log\Delta}{\varepsilon\sigma}\right) \cdot \mathrm{MST}(S) \\
&\qquad \ge (1-O(\varepsilon)) \cdot \left(\sum_{i=1}^{\log_{1+\varepsilon} W} (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)}\right) - O(\varepsilon) \cdot \mathrm{MST}(P) - O\left(\frac{\mathrm{poly}(\varepsilon)}{k \log\Delta}\right) \cdot \mathrm{MST}(S),
\end{aligned}$$

where we use $W \le O(\Delta)$, $\Gamma = \varepsilon^{-2} \cdot \log\Delta$, and the last inequality is by the definition of $\sigma = k \, \mathrm{poly}(\varepsilon^{-1} \log\Delta) \cdot \kappa$; similarly, for the other direction,

$$\begin{aligned}
\sum_{i=0}^{\log_{1+\varepsilon} W} (1+\varepsilon)^i \cdot \widetilde{c}_P^{(i)} &\le \sum_{i=0}^{\log_{1+\varepsilon} W} (1+O(\varepsilon)) \cdot (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)} + O\left(\frac{\mathrm{MST}(P)}{\Gamma}\right) \\
&\le (1+O(\varepsilon)) \left(\sum_{i=1}^{\log_{1+\varepsilon} W} (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)}\right) + O\left(\frac{\log\Delta}{\varepsilon\Gamma}\right) \cdot \mathrm{MST}(P) \\
&\le (1+O(\varepsilon)) \left(\sum_{i=1}^{\log_{1+\varepsilon} W} (1+\varepsilon)^i \cdot \mathring{c}_P^{(i)}\right) + O(\varepsilon) \cdot \mathrm{MST}(P),
\end{aligned}$$

where the last inequality is by the definition of $\Gamma = \varepsilon^{-2} \log\Delta$. By combining the above bounds and plugging in our estimate $\widetilde{\mathrm{MST}}$ in line 18 of Algorithm 6, we conclude the analysis of the error. This completes the proof of Lemma 3.7

4 Lower Bound: $\Omega(k)$ Bits are Necessary

In this section we demonstrate that any streaming algorithm for SFP achieving any finite approximation ratio for SFP requires $\Omega(k)$ bits of space.

Theorem 4.1. *For every $k > 0$, every randomized streaming algorithm achieving a finite approximation ratio for SFP with k color classes of size at most 2 must require $\Omega(k)$ bits of space. This holds even for insertion-only algorithms and even when points are from the one-dimensional line $\mathbb{R}$.*

Proof. The proof is a reduction from the INDEX problem on k bits, where Alice holds a binary string $x \in \{0,1\}^k$, and Bob has an index $i \in [k]$. The goal of Bob is to compute the bit x_i in the one-way communication model, where only Alice can send a message to Bob and not vice versa. It is well-known that Alice needs to send $\Omega(k)$ bits for Bob to succeed with constant probability [KNR99] (see also [KN97, JKS08]). Our reduction is from INDEX to SFP on the (discretized) one-dimensional line $[2k]$. Consider a randomized streaming algorithm ALG for SFP that approximates the optimal cost and in particular can distinguish whether the optimal cost is 0 or 1 with constant probability. We show that it can be used to solve the INDEX problem, implying that ALG needs to use $\Omega(k)$ bits of space.

Indeed, Alice applies ALG on the following stream: For each bit x_j, she adds to the stream a point of color j at location $2j + x_j$. So far $\text{OPT} = 0$. She now sends the internal state of ALG to Bob. Then, Bob continues the execution of ALG (using the same random coins) by adding one more point to the stream: Given his index $i \in [k]$, he adds a point of color i at location $2i$. After that, $\text{OPT} = 0 + x_i$, which is either 0 or 1. It follows that if ALG achieves a finite approximation with constant probability, then Bob can discover x_i and solve INDEX. □

5 Composing MST Sketches: k^k-time k^2-space Algorithm

As outlined in the introduction, one can solve SFP in a simple way with query time $O(k^k)\cdot\text{poly}\log\Delta$. In this section, we provide details of this approach and prove the following theorem:

Theorem 5.1. *For any integers $k, \Delta \geq 1$ and any $0 < \varepsilon < 1/2$, one can with high probability $(\alpha_2 + \varepsilon)$-approximate SFP cost of an input $X \subseteq [\Delta]^2$ presented as a dynamic geometric stream, using space and update time of $O(k^2 \cdot \text{poly}(\varepsilon^{-1} \cdot \log\Delta))$ and with query time $O(k^k)\cdot\text{poly}(\varepsilon^{-1}\cdot\log\Delta)$.*

The proof uses the streaming algorithm for MST from [FIS08] in a black box manner. Namely, we use the following result[4]:

Theorem 5.2 (Theorem 6 in [FIS08])**.** *There is an algorithm that for every $\varepsilon, \delta \in (0,1)$, integer $\Delta \geq 1$, given a (multi)set $X \subseteq [\Delta]^2$ of points presented as a dynamic geometric stream, computes a $(1+\varepsilon)$-approximate estimate for the cost of the Euclidean minimum spanning tree of X with probability at least $1-\delta$, using space $O\left(\log(1/\delta)\cdot\text{poly}(\varepsilon^{-1}\cdot\log\Delta)\right)$ and with both update and query times bounded by the same quantity. Furthermore, the algorithm returns a linear sketch from which the estimate can be computed.*

Proof of Theorem 5.1. First, we compute the MST sketch $\mathcal{K}_i$ for each color i separately, by using Theorem 5.2 with the same ε, with $\delta = 2^{-k}/3$, and also with the same random bits used (so that we are able to add up the sketches for different colors). After processing the stream and obtaining sketches $\mathcal{K}_i$, we enumerate all subsets of k colors and estimate the MST cost for all colors in the subset. Namely, for each subset $S \subseteq [k]$, we first merge (copies of) sketches $\mathcal{K}_i$ for $i \in S$, using that they are linear sketches, to get an MST sketch $\mathcal{K}_S$ for all points of colors in S. Running the estimation procedure from [FIS08] on sketch $\mathcal{K}_S$, we get an estimate of the MST cost for S and store this estimate in memory. Then, we enumerate all $O(k^k)$ partitions of k colors and for each

[4] While Theorem 6 in [FIS08] does not explicitly state that the algorithm produces a linear sketch, this follows as the data structures maintained in the algorithm are all linear sketches. We describe the MST sketch in Section 3.4.

partition $I_1, \ldots, I_r \subseteq [k]$, we estimate its cost by summing up the estimates for subsets $I_1, \ldots, I_r$, computed in the previous step. Finally, the algorithm returns the smallest estimate of a partition.

By the union bound, all estimates for subsets of $[k]$ are $(1+\varepsilon)$-approximate with probability at least $1 - \delta \cdot 2^k \geq \frac{2}{3}$, by the choice of δ. Conditioning on this, the cost of any partition is at least $1 - \varepsilon$ times the optimal cost. Consider an optimal solution F of SFP. After removing all Steiner points, we obtain an α_2-approximate solution F', by the definition of the Steiner ratio α_2. The partition of colors into components in F' is considered by the algorithm and the estimated cost of this partition is at most $(1+\varepsilon) \cdot w(F') \leq (1+\varepsilon) \cdot \alpha_2 \cdot w(F)$ with probability at least $\frac{2}{3}$, which implies the correctness of the algorithm. The space and time bounds follow from the choice of $\delta = O(2^{-k})$ and from Theorem 5.2. □

6 Future Directions

Our paper makes a progress in the understanding of geometric streaming algorithms and of applicability of Arora's framework for low-space streaming algorithms for geometric optimization problems. Still, our work leaves a number of open problems which we will discuss here.

Our approximation ratio $\alpha_2 + \varepsilon$ matches the current approximation ratio for the Steiner tree problem in geometric streams. Hence, any improvement to our approximation ratio would require to first improve the approximation for Steiner tree, even in insertion-only streams. This naturally leads to the main open problem of obtaining a $(1+\varepsilon)$-approximation for Steiner tree in geometric streams using only $\mathrm{poly}(\varepsilon^{-1} \log \Delta)$ space.

Our naïve algorithm for the Steiner forest problem given in Theorem 5.1 is also an $(\alpha_2 + \varepsilon)$-approximation with $\mathrm{poly}(k\varepsilon^{-1} \log \Delta)$ space, but its running time is exponential in k because it queries an (approximate) MST-value oracle on all possible subsets of color classes to find the minimum. We do not know if a smaller number of queries suffices here, but it is known that in a similar setup for coverage problems any oracle-based $O(1)$-approximation requires exponentially many queries to an approximate oracle [BEM17]. Thus, it would not be surprising if a similar lower bound holds for our problem.

Our Theorem 4.1 shows that for SFP with color classes of size at most 2 one cannot achieve any bounded approximation ratio using space that is sublinear in $n \leq 2k$. This strongly suggests that SFP with pairs of terminals (i.e., $C_i = \{s_i, t_i\}$) does not admit a constant-factor approximation in the streaming setting, although our lower bound construction does not extend to this case (as it requires having some size-1 color classes). We leave it as an open problem whether a constant-factor approximation in sublinear (in $n = 2k$) space is possible for this version. We notice however that for the case where both points of each terminal pair are inserted/deleted together, it is possible to get an $O(\log n)$-approximation using the metric embedding technique of Indyk [Ind04].

The main focus of this paper is on the study of SFP for the Euclidean plane, but in principle, our entire analysis can be extended to the Euclidean space $\mathbb{R}^d$, for any fixed $d \geq 2$. However, this would require extending the arguments of [BH12, BKM15], namely, the structural result that we restate in Theorem 2.2, and these details were not written explicitly in the two papers.

The techniques developed in this paper seem to be general enough to be applicable to other problems/objectives with *connectivity constraints*, where the connectivity is specified by the colors and a solution is feasible if the points of the same color are connected. One such closely related problem is the sum-of-MST objective, i.e., the problem of minimizing the sum of the costs of trees such that points of the same color are in the same tree (see also [AGLN03, ZNI05] for related problems). We hope that the approach developed in our paper can lead to a $(1+\varepsilon)$-approximation of the geometric version of this problem, using $\mathrm{poly}(k\varepsilon^{-1} \log \Delta)$ space and time (while for space only, one can use similar techniques as in Theorem 5.1). Moreover, it may be possible to apply our approach to solve the connectivity-constrained variants of other classical problems, especially those where dynamic programming has been employed successfully, like r-MST and TSP [Aro98]. For example, the TSP variant could be to find a collection of cycles of minimum total length such that points of the same color are in the same cycle.

At a higher level, the connectivity constraints may be more generally interpreted as grouping constraints. For instance, in the context of clustering, our color constraints may be viewed as *must-link* constraints, where points of the same color have to be placed in the same cluster. Such constrained clustering framework is of significant interest in data analysis (see, e.g., a highly-cited paper [WCRS01]). Our framework, combined with coreset techniques [FS05] and Arora's quad-tree methods (see [ARR98]), may be used to design streaming algorithms for such clustering problems.

Finally, we believe that the framework of optimization problems with connectivity and grouping constraints is interesting on its own, going beyond the streaming setup. Such problems may be studied also in the setting of standard (offline) algorithms, as well as of online algorithms, approximation algorithms, fixed-parameter tractability, and heuristics.

References

[ABIW09] Alexandr Andoni, Khanh Do Ba, Piotr Indyk, and David P. Woodruff. Efficient sketches for earth-mover distance, with applications. In *Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 324–330, 2009.

[AGLN03] Mattias Andersson, Joachim Gudmundsson, Christos Levcopoulos, and Giri Narasimhan. Balanced partition of minimum spanning trees. *International Journal of Computational Geometry and Applications*, 13(4):303–316, 2003.

[AIK08] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *Proceedings of the 19th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 343–352, 2008.

[AKR95] Ajit Agrawal, Philip N. Klein, and R. Ravi. When trees collide: An approximation algorithm for the generalized Steiner problem on networks. *SIAM Journal on Computing*, 24(3):440–456, 1995.

[AN12] Alexandr Andoni and Huy L. Nguyen. Width of points in the streaming model. In *Proceedings of the 23rd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 447–452, 2012.

[ANOY14] Alexandr Andoni, Aleksandar Nikolov, Krzysztof Onak, and Grigory Yaroslavtsev. Parallel algorithms for geometric graph problems. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing (STOC)*, pages 574–583, 2014.

[Aro98] Sanjeev Arora. Polynomial time approximation schemes for Euclidean traveling salesman and other geometric problems. *Journal of the ACM*, 45(5):753–782, 1998.

[ARR98] Sanjeev Arora, Prabhakar Raghavan, and Satish Rao. Approximation schemes for Euclidean k-medians and related problems. In *Proceedings of the 13th Annual ACM Symposium on the Theory of Computing (STOC)*, pages 106–113, 1998.

[Bar96] Yair Bartal. Probabilistic approximations of metric spaces and its algorithmic applications. In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 184–193, 1996.

[BEM17] MohammadHossein Bateni, Hossein Esfandiari, and Vahab S. Mirrokni. Almost optimal streaming algorithms for coverage problems. In *Proceedings of the 29th ACM Symposium on Parallelism in Algorithms and Architectures (SPAA)*, pages 13–23, 2017.

[BFL$^+$17] Vladimir Braverman, Gereon Frahling, Harry Lang, Christian Sohler, and Lin F. Yang. Clustering high dimensional dynamic data streams. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 576–585, 2017.

[BH12] MohammadHossein Bateni and MohammadTaghi Hajiaghayi. Euclidean prize-collecting Steiner forest. *Algorithmica*, 62(3-4):906–929, 2012.

[BHM11] MohammadHossein Bateni, MohammadTaghi Hajiaghayi, and Dániel Marx. Approximation schemes for Steiner forest on planar graphs and graphs of bounded treewidth. *Journal of the ACM*, 58(5):21:1–21:37, 2011.

[BKM15] Glencora Borradaile, Philip N. Klein, and Claire Mathieu. A polynomial-time approximation scheme for Euclidean Steiner forest. *ACM Transactions of Algorithms*, 11(3):19:1–19:20, 2015.

[BZ16] Djamal Belazzougui and Qin Zhang. Edit distance: Sketching, streaming, and document exchange. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 51–60, 2016.

[CF14] Graham Cormode and Donatella Firmani. A unifying framework for ℓ_0-sampling algorithms. *Distributed Parallel Databases*, 32(3):315–335, 2014.

[CFH$^+$21] Kuan Cheng, Alireza Farhadi, MohammadTaghi Hajiaghayi, Zhengzhong Jin, Xin Li, Aviad Rubinstein, Saeed Seddighin, and Yu Zheng. Streaming and small space approximation algorithms for edit distance and longest common subsequence. In *Proceedings of the 48th International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 54:1–54:20, 2021.

[CG85] F. R. K. Chung and R. L. Graham. A new bound for Euclidean Steiner minimal trees. *Annals of the New York Academy of Sciences*, 440(1):328–346, 1985.

[CGK16] Diptarka Chakraborty, Elazar Goldenberg, and Michal Koucký. Streaming algorithms for embedding and computing edit distance in the low distance regime. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, pages 712–725, 2016.

[Cha02] M. Charikar. Similarity estimation techniques from rounding algorithms. In *34th annual ACM Symposium on Theory of Computing*, pages 380–388. ACM Press, 2002. doi:http://doi.acm.org/10.1145/509907.509965.

[Cha06] Timothy M. Chan. Faster core-set constructions and data-stream algorithms in fixed dimensions. *Computation Geometry*, 35(1-2):20–35, 2006.

[Cha16] Timothy M. Chan. Dynamic streaming algorithms for ε-kernels. In *Proceedings of the 32nd International Symposium on Computational Geometry (SoCG)*, pages 27:1–27:11, 2016.

[CHJ18] T.-H. Hubert Chan, Shuguang Hu, and Shaofeng H.-C. Jiang. A PTAS for the Steiner forest problem in doubling metrics. *SIAM Journal on Computing*, 47(4):1705–1734, 2018.

[CLMS13] Artur Czumaj, Christiane Lammersen, Morteza Monemizadeh, and Christian Sohler. $(1+\varepsilon)$-approximation for facility location in data streams. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1710–1728, 2013.

[CM06] Graham Cormode and S. Muthukrishnan. Combinatorial algorithms for compressed sensing. In *Proceedings of the 13th International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, pages 280–294, 2006.

[CRT05] Bernard Chazelle, Ronitt Rubinfeld, and Luca Trevisan. Approximating the minimum spanning tree weight in sublinear time. *SIAM Journal on Computing*, 34(6):1370–1379, 2005.

[EJ15] Funda Ergün and Hossein Jowhari. On the monotonicity of a data stream. *Combinatorica*, 35(6):641–653, 2015.

[FIS08] Gereon Frahling, Piotr Indyk, and Christian Sohler. Sampling in dynamic data streams and applications. *International Journal of Computational Geometry and Applications*, 18(1/2):3–28, 2008.

[FKZ05] Joan Feigenbaum, Sampath Kannan, and Jian Zhang. Computing diameter in the streaming and sliding-window models. *Algorithmica*, 41(1):25–41, 2005.

[FS05] Gereon Frahling and Christian Sohler. Coresets in dynamic geometric data streams. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 209–217, 2005.

[GGK$^+$18] Martin Groß, Anupam Gupta, Amit Kumar, Jannik Matuschke, Daniel R. Schmidt, Melanie Schmidt, and José Verschae. A local-search algorithm for Steiner forest. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, pages 31:1–31:17, 2018.

[GJKK07] Parikshit Gopalan, T. S. Jayram, Robert Krauthgamer, and Ravi Kumar. Estimating the sortedness of a data stream. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 318–327, 2007.

[GK15] Anupam Gupta and Amit Kumar. Greedy algorithms for Steiner forest. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 871–878, 2015.

[GP68] Edgar N. Gilbert and Henry O. Pollak. Steiner minimal trees. *SIAM Journal on Applied Mathematics*, 16(1):1–29, 1968.

[GW95] Michel X. Goemans and David P. Williamson. A general approximation technique for constrained forest problems. *SIAM Journal on Computing*, 24(2):296–317, 1995.

[HM04] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 291–300, 2004.

[HSYZ19] Wei Hu, Zhao Song, Lin F. Yang, and Peilin Zhong. Nearly optimal dynamic k-means clustering for high-dimensional data, 2019. arXiv:1802.00459.

[Ind04] Piotr Indyk. Algorithms for dynamic geometric problems over data streams. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 373–380, 2004.

[IT03] Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *ICCV*, 2003.

[Jai01] Kamal Jain. A factor 2 approximation algorithm for the generalized Steiner network problem. *Combinatorica*, 21(1):39–60, 2001.

[JKS08] T. S. Jayram, Ravi Kumar, and D. Sivakumar. The one-way communication complexity of Hamming distance. *Theory of Computing*, 4(6):129–135, 2008.

[KN97] Eyal Kushilevitz and Noam Nisan. *Communication Complexity*. Cambridge University Press, 1997.

[KNPW11] Daniel M. Kane, Jelani Nelson, Ely Porat, and David P. Woodruff. Fast moment estimation in data streams in optimal space. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC)*, pages 745–754, 2011.

[KNR99] Ilan Kremer, Noam Nisan, and Dana Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

[KNW10] Daniel M. Kane, Jelani Nelson, and David P. Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings of the 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*, pages 41–52, 2010.

[LNNT19] Kasper Green Larsen, Jelani Nelson, Huy L. Nguyen, and Mikkel Thorup. Heavy hitters via cluster-preserving clustering. *Communications of the ACM*, 62(8):95–100, 2019.

[LS08] Christiane Lammersen and Christian Sohler. Facility location in dynamic geometric data streams. In *Proceedings of the 16th Annual European Symposium on Algorithms (ESA)*, pages 660–671, 2008.

[Mit99] Joseph S. B. Mitchell. Guillotine subdivisions approximate polygonal subdivisions: A simple polynomial-time approximation scheme for geometric TSP, k-MST, and related problems. *SIAM Journal on Computing*, 28(4):1298–1309, 1999.

[MW95] Thomas L. Magnanti and Laurence A. Wolsey. Chapter 9: Optimal trees. In *Network Models*, volume 7 of *Handbooks in Operations Research and Management Science*, pages 503–615. Elsevier, 1995.

[Pol90] David Pollard. *Empirical Processes: Theory and Applications*, chapter 4: Packing and Covering in Euclidean Spaces, pages 14–20. IMS, 1990.

[Sch16] Guido Schäfer. Steiner Forest. In Ming-Yang Kao, editor, *Encyclopedia of Algorithms*, pages 2099–2102. Springer, New York, NY, 2016.

[Soh12] Christian Sohler. Problem 52: TSP in the streaming model. https://sublinear.info/52, 2012.

[SS13] Michael E. Saks and C. Seshadhri. Space efficient streaming algorithms for the distance to monotonicity and asymmetric edit distance. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1698–1709, 2013.

[SW07] Xiaoming Sun and David P. Woodruff. The communication and streaming complexity of computing the longest common and increasing subsequences. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 336–345, 2007.

[WCRS01] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 577–584, 2001.

[ZNI05] Liang Zhao, Hiroshi Nagamochi, and Toshihide Ibaraki. Greedy splitting algorithms for approximating multiway partition problems. *Mathematical Programming, Series A*, 102(1):167–183, 2005.

A Technical Lemmas

Lemma A.1. *There exists an algorithm that, given an integer n, a stream of dynamic updates to a frequency vector $U \in [-n, n]^n$, and an (integer) threshold $T \geq 1$, with probability at least $1 - 1/\operatorname{poly}(n)$ it reports YES if $0 < \|U\|_0 \leq T$ and NO if $\|U\|_0 > 2T$ (otherwise, the answer may be arbitrary), and if the answer is YES, it returns all the non-zero coordinates of U, using space $O(T \cdot \operatorname{poly}\log n)$ and the same time per update.*

Proof. The algorithm maintains an ℓ_0-norm estimator for U (see e.g. [KNW10]) with relative error $\varepsilon = 0.5$ using space $\operatorname{poly}\log n$, and a compressed sensing structure that recovers $2T$ non-zero elements of U (see e.g. [CM06]) using space $O(T \operatorname{poly}\log n)$, each succeeding with probability at least $1 - 1/\operatorname{poly}(n)$. When the stream ends, the algorithm queries the ℓ_0-norm estimator, which is accurate enough to distinguish whether $\|U\|_0 \leq T$ or $\|U\|_0 > 2T$. conditioned on this estimator succeeding, if it is determined that $\|U\|_0 \leq 2T$, the algorithm uses the compressed sensing structure to recover the at most $2T$ non-zero coordinates of U. This finishes the proof of Lemma A.1. □

We remark that an alternative algorithm is to maintain $O(T \cdot \operatorname{poly}\log n)$ independent instances of an ℓ_0-sampler (cf. Lemma 3.10) and applying a coupon collector argument to both estimate the ℓ_0 norm and recover the non-zero coordinates. The connection goes also the other way: An ℓ_0-sampler may be designed using this lemma together with an appropriate subsampling of the domain.