Spanning trees with $O(1)$ average stretch factor

Michiel Smid[*]

April 28, 2009

Abstract

Let G be a connected graph with n vertices in which each edge has a weight, and let T be a spanning tree of G. The stretch factor of two vertices x and y is the ratio of the distance between x and y in T and the shortest-path distance between x and y in G. In SODA 2007, Abraham, Bartal and Neiman showed that there exists a spanning tree T of G such that the average stretch factor (over all $\binom{n}{2}$ vertex pairs) is bounded by a constant.

We prove this result for the cases when (i) G is the complete graph on a set of points in $\mathbb{R}^d$ and edge weights represent Euclidean distances and (ii) G is the complete graph on a set of points in a metric space and edge weights represent distances in this space.

1 Introduction

Let $(S, \mathbf{d})$ be a finite metric space and let H be a connected edge-weighted graph with vertex set S in which the weight of any edge (x, y) is equal to $\mathbf{d}(x, y)$. The length of a path in H is defined to be the sum of the weights of the edges on the path. For any two points x and y of S, we denote by $\mathbf{d}_H(x, y)$ the minimum length of any path in H between x and y. If $x \neq y$, then the *stretch factor* of x and y is defined to be $\mathbf{d}_H(x, y)/\mathbf{d}(x, y)$. If $t \geq 1$ is a real number such that each pair of distinct points in S has stretch factor

[*]School of Computer Science, Carleton University, Ottawa, Ontario, Canada K1S 5B6.
Research supported by NSERC.

at most t, then we say that H is a *t-spanner* of S. The smallest value of t such that H is a t-spanner of S is called the *stretch factor* of H. (For an overview of results on t-spanners for the Euclidean metric, see the book by Narasimhan and Smid [8].) Observe that any t-spanner of S must have at least $n - 1$ edges.

Assume that S is a set of n points in $\mathbb{R}^d$ (where the dimension d is a constant) and $\mathbf{d}$ is the Euclidean distance function. Das and Heffernan [5] have shown that for any real constant $\epsilon > 0$, there exists a graph H with at most $(1 + \epsilon)n$ edges, such that H is a t-spanner of S, for some constant t that depends on ϵ and d. Aronov *et al.* [3] have shown that this result is optimal: For any constant $t > 1$, t-spanners with $n + o(n)$ edges do not exist for all sets of n points in $\mathbb{R}^d$.

For an arbitrary finite metric space $(S, \mathbf{d})$, with $|S| = n$, it is not difficult to show that a minimum spanning tree is an $(n-1)$-spanner of S. Eppstein [7] has shown that this result cannot be improved: If S is the vertex set of a regular n-gon in the plane and $\mathbf{d}$ is the Euclidean distance function, then any spanning tree of S has stretch factor $\Omega(n)$. (An alternative proof of this lower bound is given in [3].)

In this paper, we consider the problem of constructing a graph H whose *average* stretch factor[1]

$$ASF(H) = \frac{1}{\binom{n}{2}} \sum_{\{x,y\} \in \mathcal{P}_2(S)} \frac{\mathbf{d}_H(x, y)}{\mathbf{d}(x, y)}$$

is bounded by a constant and that contains as few edges as possible. Since H must be a connected graph, it contains at least $n - 1$ edges.

The result of Das and Heffernan implies that, for the Euclidean metric and for any real constant $\epsilon > 0$, there exists a graph H having at most $(1 + \epsilon)n$ edges, such that $ASF(H) = O(1)$.

Consider Eppstein's example of the vertex set S of a regular n-gon. Even though any spanning tree of S has stretch factor $\Omega(n)$, there exists a spanning tree T whose average stretch factor $ASF(T)$ is bounded from above by a constant: If we take for T the tree (in fact, the path) obtained by deleting a random edge of the n-gon, then the expected value of $ASF(T)$ is bounded by a constant. (A proof of this claim follows from results by Alon *et al.* [2]. In fact, this result holds for any set of points on a circle.)

[1] $\mathcal{P}_2(S)$ denotes the set of all $\binom{n}{2}$ unordered pairs of distinct elements in S.

Abraham *et al.* [1] have shown that a spanning tree T with $ASF(T) = O(1)$ exists for any finite metric space[2]:

Theorem 1 *There exists a constant $\alpha > 1$, such that every finite metric space $(S, \mathbf{d})$ contains a spanning tree T such that $ASF(T) \leq \alpha$. Such a spanning tree can be computed in polynomial time.*

In this note, we present an alternative proof of Theorem 1, which is simpler to understand than the proof in [1]. (Of course, the reason that our proof is simpler is the fact that the result in [1] is stronger.)

In Section 2, we prove Theorem 1 for the case when S is a set of n points in $\mathbb{R}^d$ and $\mathbf{d}$ is the Euclidean distance function. In this case, the spanning tree is obtained from Callahan and Kosaraju's split tree (see [4]), where the splitting of the bounding box of the point set is done in a "careful" way. In Section 4, we prove Theorem 1 for arbitrary metric spaces (and, in fact, obtain a better value for the constant α).

Our construction uses the following lemma, which states that any sequence of real numbers can be cut in the "middle third", such that any pair of elements in the sequence that are very close together are on opposite sides of the cut:

Lemma 1 *There exists a constant $\beta > 2$ such that the following is true. Let $n \geq 2$ be an integer and let $x_1 \leq x_2 \leq \cdots \leq x_n$ be a sequence of real numbers with $x_1 \neq x_n$. Then, there exists a real number z such that*

$$x_1 + \frac{x_n - x_1}{3} \leq z \leq x_1 + \frac{2(x_n - x_1)}{3}$$

and

$$\sum_{x_i \leq z} \sum_{x_j > z} \frac{1}{x_j - x_i} \leq \frac{\beta}{x_n - x_1} m(n-m),$$

where $m = |\{i : x_i \leq z\}|$.

We remark that a simple probabilistic argument shows that such a z with

$$\sum_{x_i \leq z} \sum_{x_j > z} \frac{1}{x_j - x_i} \leq \frac{3}{x_n - x_1} \binom{n}{2}$$

[2] In fact, they prove that every weighted graph contains such a spanning tree.

exists; see Lemma 4. Lemma 1, however, states that $\binom{n}{2}$ can be replaced by the *smaller* value $m(n-m)$, which counts the number of pairs of elements that are in different subsequences of the partition. The proof of Lemma 1 will be given in Section 3.

2 The Euclidean metric

Throughout this section, S denotes a finite set of points in $\mathbb{R}^d$ and $\mathbf{d}$ denotes the Euclidean distance function.

A *hyperrectangle* is defined to be the Cartesian product of d closed intervals. Hence, such a hyperrectangle R can be written as

$$R = [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_d, b_d],$$

where a_i and b_i are real numbers with $a_i \leq b_i$, $1 \leq i \leq d$. We call $L_i(R) = b_i - a_i$ the *side length* of R along the i-th dimension. We define $L_{\max}(R)$ to be the maximum side length of R along any dimension. The *bounding box* of the point set S is defined to be the smallest hyperrectangle that contains all points of S.

The algorithm that computes a spanning tree of S is as follows:

Algorithm EuclLowAverStrTree(S)
Input: A finite set S of points in $\mathbb{R}^d$.
Output: A pair (T, r), where T is a spanning tree of S and r is the root of T.

if $|S| = 1$
  then let p be the element of S;
    let T be the tree consisting of the single node p;
    return (T, p);
  else let R be the bounding box of S;
    let i be the dimension such that $L_{\max}(R) = L_i(R)$;
    let $x_1 \leq x_2 \leq \ldots \leq x_n$ denote the sorted sequence of the i-th coordinates of the points in S;
    let z be a real number as given by Lemma 1;
    let S_1 be the set of all points of S whose i-th coordinates are at most z;
    let $S_2 = S \setminus S_1$;
    $(T_1, r_1) =$ EuclLowAverStrTree(S_1);

```
12.        (T_2, r_2) = EuclLowAverStrTree(S_2);
13.        let T be the union of T_1, T_2 and the edge (r_1, r_2);
14.        return (T, r_1)
```

In Lemma 3 below, we will prove that the average stretch factor of the spanning tree T that is returned by this algorithm is bounded by a constant. Before we can prove this claim, we show that the length of any path in T from the root to any point q has length $O(L_{\max}(R))$:

Lemma 2 *Let R be the bounding box of S, let T be the spanning tree of S that is returned by algorithm* EuclLowAverStrTree(S)*, and let p be the root of T. Then, for any point q in S, we have*

$$\mathbf{d}_T(p,q) \leq 3d\sqrt{d} \cdot L_{\max}(R).$$

Proof. If $|S| = 1$, then $q = p$ and $L_{\max}(R) = 0$, and, therefore, the lemma obviously holds. Assume that $|S| \geq 2$. Let $L = L_{\max}(R)$, let i be the dimension such that $L_i(R) = L$, and consider the sets S_1 and S_2 that are computed in lines 9 and 10 of the algorithm.

First observe that the diameter of S is at most $\sqrt{d}L$. Therefore, the distance between the roots of the recursively computed trees T_1 and T_2 is at most $\sqrt{d}L$. Next, it follows from Lemma 1 that the side lengths along the i-th dimension of the bounding boxes of S_1 and S_2 are at most $2L/3$. Finally, after d recursive calls, all side lengths of the bounding box of the current point set are at most $2L/3$. It follows that

$$\mathbf{d}_T(p,q) \leq d\sqrt{d}L \sum_{j=0}^{\infty} (2/3)^j = 3d\sqrt{d}L.$$

∎

Lemma 3 *Assume that $|S| \geq 2$ and let T be the spanning tree of S that is returned by algorithm* EuclLowAverStrTree(S)*. Then*

$$ASF(T) \leq 6\beta d\sqrt{d},$$

where β is the constant in Lemma 1.

Proof. We will prove by induction on the size of S that

$$\sum_{\{x,y\}\in\mathcal{P}_2(S)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} \leq 6\beta d\sqrt{d}\binom{n}{2}.$$

If S contains only one element, then T is the tree consisting of one single node. In this case, both the lefthand and righthand sides are equal to zero.

Assume that S contains at least two elements. Let R be the bounding box of S, let $L = L_{\max}(R)$, and consider the sets S_1 and S_2 that are computed in lines 9 and 10 of the algorithm.

The tree T is the union of (i) the recursively computed spanning tree T_1 of S_1, (ii) the recursively computed spanning tree T_2 of S_2, and (iii) the edge (r_1, r_2) joining the roots of T_1 and T_2. Let $p = r_1$; thus, p is the root of T. We have

$$\sum_{\{x,y\}\in\mathcal{P}_2(S)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} = \sum_{j=1}^{2} \sum_{\{x,y\}\in\mathcal{P}_2(S_j)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} + \sum_{x\in S_1, y\in S_2} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)}.$$

If both x and y are in the same subset S_j, then $\mathbf{d}_T(x,y) = \mathbf{d}_{T_j}(x,y)$. Thus, by induction, we have

$$\sum_{\{x,y\}\in\mathcal{P}_2(S_j)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} = \sum_{\{x,y\}\in\mathcal{P}_2(S_j)} \frac{\mathbf{d}_{T_j}(x,y)}{\mathbf{d}(x,y)} \leq 6\beta d\sqrt{d}\binom{|S_j|}{2}.$$

For any point x in S_1 and any point y in S_2, we have, by Lemma 2,

$$\mathbf{d}_T(x,y) = \mathbf{d}_T(x,p) + \mathbf{d}_T(p,y) \leq 6d\sqrt{d}L.$$

It follows that

$$\sum_{x\in S_1, y\in S_2} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} \leq 6d\sqrt{d}L \sum_{x\in S_1, y\in S_2} \frac{1}{\mathbf{d}(x,y)}.$$

Let i be the dimension such that $L_i(R) = L$. Observe that, for $x \in S_1$ and $y \in S_2$, the Euclidean distance $\mathbf{d}(x,y)$ is at least the difference between the i-th coordinates of the points y and x. Therefore, the choice of z in line 8 of the algorithm and Lemma 1 imply that

$$\sum_{x\in S_1, y\in S_2} \frac{1}{\mathbf{d}(x,y)} \leq \frac{\beta}{L}|S_1| \cdot |S_2|.$$

Thus, we have

$$\sum_{x \in S_1, y \in S_2} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} \le 6\beta d\sqrt{d}\, |S_1| \cdot |S_2|.$$

It follows that

$$\begin{aligned}\sum_{\{x,y\} \in \mathcal{P}_2(S)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} &\le 6\beta d\sqrt{d} \left(\binom{|S_1|}{2} + \binom{|S_2|}{2} + |S_1| \cdot |S_2| \right) \\ &= 6\beta d\sqrt{d} \binom{n}{2}.\end{aligned}$$

∎

3 The proof of Lemma 1

Our proof of Lemma 1 uses the following weaker lemma:

Lemma 4 *Let $n \ge 2$ be an integer, let $x_1 \le x_2 \le \cdots \le x_n$ be a sequence of real numbers with $x_1 \ne x_n$, and let a and b be two real numbers such that $x_1 \le a < b \le x_n$. Then, there exists a real number z such that $a < z < b$ and*

$$\sum_{x_i \le z} \sum_{x_j > z} \frac{1}{x_j - x_i} \le \frac{1}{b-a} \binom{n}{2}.$$

Proof. Let z be a real number that is chosen uniformly at random in the interval (a,b). For any two indices i and j with $1 \le i < j \le n$, define the indicator random variable

$$X_{ij} = \begin{cases} 1 & \text{if } x_i \le z < x_j, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\sum_{x_i \le z} \sum_{x_j > z} \frac{1}{x_j - x_i} = \sum_{1 \le i < j \le n} X_{ij} \cdot \frac{1}{x_j - x_i}$$

and, by the linearity of expectation,

$$E\left(\sum_{x_i \le z} \sum_{x_j > z} \frac{1}{x_j - x_i} \right) = \sum_{1 \le i < j \le n} E(X_{ij}) \cdot \frac{1}{x_j - x_i}.$$

Consider two indices i and j with $1 \le i < j \le n$. We prove an upper bound on the expected value $E(X_{ij})$ of the random variable X_{ij}. If $a \le x_i < x_j \le b$, then

$$E(X_{ij}) = \Pr(X_{ij} = 1) = \Pr(x_i \le z < x_j) = \frac{x_j - x_i}{b - a}.$$

If $x_i < a \le x_j \le b$. Then

$$E(X_{ij}) = \Pr(X_{ij} = 1) = \Pr(a < z < x_j) = \frac{x_j - a}{b - a} \le \frac{x_j - x_i}{b - a}.$$

If $a \le x_i \le b < x_j$, then

$$E(X_{ij}) = \Pr(X_{ij} = 1) = \Pr(x_i \le z < b) = \frac{b - x_i}{b - a} \le \frac{x_j - x_i}{b - a}.$$

If $x_i \le a$ and $x_j \ge b$, then $X_{ij} = 1$ and

$$E(X_{ij}) = 1 \le \frac{x_j - x_i}{b - a}.$$

In all other cases, we have $X_{ij} = 0$ and

$$E(X_{ij}) = 0 \le \frac{x_j - x_i}{b - a}.$$

It follows that

$$E\left(\sum_{x_i \le z} \sum_{x_j > z} \frac{1}{x_j - x_i}\right) \le \sum_{1 \le i < j \le n} \frac{1}{b - a} = \frac{1}{b - a}\binom{n}{2}.$$

Thus, there exists a real number z such that $a < z < b$ and

$$\sum_{x_i \le z} \sum_{x_j > z} \frac{1}{x_j - x_i} \le \frac{1}{b - a}\binom{n}{2}.$$

∎

We now prove Lemma 1. Let $n \ge 2$ be an integer and let $x_1 \le x_2 \le \cdots \le x_n$ be a sequence of real numbers with $x_1 \ne x_n$. We have to show that there exists a real number z in the middle third of the interval $[x_1, x_n]$ such that

$$\sum_{x_i \le z} \sum_{x_j > z} \frac{1}{x_j - x_i} \le \frac{\beta}{x_n - x_1}\, m(n - m), \tag{1}$$

where $m = |\{i : x_i \leq z\}|$ and $\beta > 2$ is a constant.

We will prove this claim by induction on n. During the induction proof, we will determine the value of β.

If $n = 2$, then we take $z = (x_1 + x_2)/2$. In this case, $m = 1$ and (1) obviously holds. Let $n \geq 3$ and assume that the claim holds for all sequences having less than n elements. Let $L = x_n - x_1$. We may assume without loss of generality that $x_1 = 0$ and $x_n = L$. Define the intervals $I_0 = [0, L/9]$ and, for each integer i with $1 \leq i \leq 8$, $I_i = (iL/9, (i+1)L/9]$. Observe that these nine intervals are pairwise disjoint and cover the interval $[x_1, x_n]$.

Let δ be a real number with $0 < \delta < 1/9$. We say that an interval I_i is *heavy*, if it contains at least δn elements of the sequence $x_1, x_2, \ldots, x_n$. If I_i contains less than δn elements, then we say that this interval is *light*. Since $\delta < 1/9$, there is at least one heavy interval.

Let k be the smallest index such that the interval I_k is heavy, and let ℓ be the largest index such that the interval I_ℓ is heavy. Observe that $k \leq \ell$. We distinguish two cases.

Case 1: There exists an index k' such that $3 \leq k' \leq 5$ and $k < k' < \ell$.

By Lemma 4, there exists a real number $z \in I_{k'}$ such that

$$\sum_{x_i \leq z} \sum_{x_j > z} \frac{1}{x_j - x_i} \leq \frac{9}{L} \binom{n}{2} \leq \frac{9}{2L} n^2.$$

Observe that z is in the middle third of the interval $[x_1, x_n]$. Let $m = |\{i : x_i \leq z\}|$. Since I_k is heavy and all elements in I_k are less than z, we have $m \geq \delta n$. Similarly, since I_ℓ is heavy and all elements in I_ℓ are larger than z, we have $n - m \geq \delta n$. It follows that

$$\sum_{x_i \leq z} \sum_{x_j > z} \frac{1}{x_j - x_i} \leq \frac{9}{2L}\, n \cdot n \leq \frac{9}{2\delta^2 L}\, m(n-m).$$

Thus, if we choose β such that

$$\beta \geq \frac{9}{2\delta^2}, \tag{2}$$

then (1) holds.

Case 2: An index k' as in Case 1 does not exist.

We claim that (i) both k and ℓ are at least 4 or (ii) both k and ℓ are at most 4. To prove this, recall that $k \leq \ell$. Thus, if $k \geq 4$, then (i) holds. If

$k \leq 3$, then $\ell \leq 4$ (because, otherwise, we can take $k' = 4$ and are in Case 1) and, therefore, (ii) holds.

We may assume without loss of generality that (i) holds. Thus, $\ell \geq k \geq 4$ and each of the intervals I_0, I_1, I_2, and I_3 is light.

First assume that $I_3 \cap \{x_1, \ldots, x_n\} = \emptyset$. Let z be an arbitrary real number in I_3 and let $m = |\{i : x_i \leq z\}|$. Then z is in the middle third of the interval $[x_1, x_n]$. Since the length of the interval I_3 is equal to $L/9$, we have

$$\sum_{x_i \leq z} \sum_{x_j > z} \frac{1}{x_j - x_i} \leq \frac{9}{L} m(n-m).$$

Thus, since β satisfies the condition in (2) and $\beta < 1/9$, it follows that (1) holds.

From now on, we assume that $I_3 \cap \{x_1, \ldots, x_n\} \neq \emptyset$. Let a and b be the indices such that x_a and x_b are the minimum and maximum elements in $I_3 \cap \{x_1, \ldots, x_n\}$, respectively.

If $x_a > 10L/27$, i.e., x_a is not in the left third of the interval I_3, then we take for z an arbitrary real number with $L/3 < z < 10L/27$. Thus, z is in the left third of I_3. In this case, letting $m = |\{i : x_i \leq z\}|$, we have

$$\sum_{x_i \leq z} \sum_{x_j > z} \frac{1}{x_j - x_i} \leq \frac{27}{L} m(n-m).$$

Since β satisfies the condition in (2) and $\beta < 1/9$, it follows that (1) holds.

By a symmetric argument, if $x_b < 11L/27$, i.e., x_b is not in the right third of I_3, then we take for z an arbitrary real number with $11L/27 < z < 4L/9$. Thus, z is in the right third of I_3. In this case, (1) holds.

Thus, we may assume that x_a is in the left third of I_3 and x_b is in the right third of I_3. Consider the sequence $x_a, x_{a+1}, \ldots, x_b$, and let $L' = x_b - x_a$. Observe that $L' \geq L/27$. By the induction hypothesis, there exists a real number z in the middle third of the interval $[x_a, x_b]$ such that

$$\sum_{x_a \leq x_i \leq z} \sum_{z < x_j \leq x_b} \frac{1}{x_j - x_i} \leq \frac{\beta}{L'} m'(b-a+1-m'),$$

where $m' = |\{i : a \leq i \leq b \text{ and } x_i \leq z\}|$. Define $m = |\{i : 1 \leq i \leq n \text{ and } x_i \leq z\}|$. Observe that $m' \leq m$. Since each of the intervals $I_0, I_1, I_2,$

and I_3 is light, we have $m \leq 4\delta n$ and, therefore, $n - m \geq (1 - 4\delta)n$. Furthermore, since I_3 is light, we have $b - a + 1 - m' \leq \delta n$. Thus,

$$b - a + 1 - m' \leq \delta n \leq \frac{\delta}{1 - 4\delta}\,(n - m).$$

It follows that

$$\sum_{x_a \leq x_i \leq z} \sum_{z < x_j \leq x_b} \frac{1}{x_j - x_i} \leq \frac{27\delta\beta}{(1 - 4\delta)L}\, m(n - m).$$

Recall that z is in the middle third of the interval $[x_a, x_b]$. If (i) $x_i < x_a$ and $x_j > z$ or (ii) $x_a \leq x_i \leq z$ and $x_j > x_b$, then $x_j - x_i > (x_b - x_a)/3 = L'/3 \geq L/81$. The number of pairs (x_i, x_j) for which (i) or (ii) holds is at most $m(n - m)$. It follows that

$$\sum_{x_i < x_a} \sum_{x_j > z} \frac{1}{x_j - x_i} + \sum_{x_a \leq x_i \leq z} \sum_{x_j > x_b} \frac{1}{x_j - x_i} \leq \frac{81}{L}\, m(n - m).$$

By combining the above inequalities, we obtain

$$\begin{aligned}
&\sum_{x_i \leq z} \sum_{x_j > z} \frac{1}{x_j - x_i} \\
&\quad = \sum_{x_i < x_a} \sum_{x_j > z} \frac{1}{x_j - x_i} + \sum_{x_a \leq x_i \leq z} \sum_{z < x_j \leq x_b} \frac{1}{x_j - x_i} + \sum_{x_a \leq x_i \leq z} \sum_{x_j > x_b} \frac{1}{x_j - x_i} \\
&\quad \leq \frac{81}{L}\, m(n - m) + \frac{27\delta\beta}{(1 - 4\delta)L}\, m(n - m).
\end{aligned}$$

The quantity on the right-hand side is less than or equal to $\frac{\beta}{L}\, m(n - m)$ if and only if

$$81 + \frac{27\delta\beta}{1 - 4\delta} \leq \beta.$$

Thus, if we choose β such that

$$\beta \geq \frac{81(1 - 4\delta)}{1 - 31\delta}, \tag{3}$$

and choose δ such that $\delta < 1/31$, then (1) holds.

If we take $\delta = 1/32$ and $\beta = 4608$, then the requirements in (2) and (3) are satisfied. Thus, we have shown that Lemma 1 holds with $\beta = 4608$.

4 General metric spaces

In this section, we generalize the construction of Section 2 to arbitrary metric spaces. Our proof of Theorem 1 is based on a slightly modified version of the low-cost star decomposition of Elkin *et al.* [6].

Let $(S, \mathbf{d})$ be a finite metric space. For any point p in S and any real number $r \geq 0$, the *ball* with center p and radius r is defined to be the set $\{x \in S : \mathbf{d}(p,x) \leq r\}$. For any subset X of S and for any point p in X, we define the *radius* $rad_X(p)$ of X with respect to p as

$$rad_X(p) = \max\{\mathbf{d}(p,x) : x \in X\}.$$

Thus, $rad_X(p)$ is the minimum radius of any ball centered at p that contains all points of X.

Consider a partition of the set S into subsets $S_1, S_2, \ldots, S_k$, for some $k \geq 2$. We define $D(S_1, \ldots, S_k)$ to be the set of all (unordered) pairs $\{x, y\}$ in $\mathcal{P}_2(S)$ that are in different subsets of the partition. Thus,

$$D(S_1, \ldots, S_k) = \bigcup_{i=1}^{k-1} \bigcup_{j=i+1}^{k} \{\{x,y\} : x \in S_i, y \in S_j\}.$$

The following lemma, which forms the basis of our algorithm for computing a spanning tree with low average stretch factor, states the following: There exists a partition of S into subsets $S_1, S_2, \ldots, S_k$, such that (i) the radius of each subset is at most a constant factor of the radius of S and (ii) points x and y whose distance is very small are in the same subset of the partition.

Lemma 5 *There exists a constant $\gamma > 2$ such that the following is true. Let p be an arbitrary element of S. There exists a partition $S_1, S_2, \ldots, S_k$ of S, for some $k \geq 2$, and a sequence $p_1, p_2, \ldots, p_k$ of points in S, such that*

1. *$p_1 = p$,*
2. *$p_i \in S_i$ and $rad_{S_i}(p_i) \leq \frac{2}{3} \cdot rad_S(p)$ for all i with $1 \leq i \leq k$, and*
3. *$\sum_{\{x,y\} \in D} \frac{1}{\mathbf{d}(x,y)} \leq \frac{\gamma}{rad_S(p)} |D|$, where $D = D(S_1, \ldots, D_k)$.*

The proof of this lemma will be given in Section 5. The algorithm that computes a spanning tree of S is as follows:

Algorithm LowAverStrTree$(S, \mathbf{d}, p)$
Input: A finite metric space $(S, \mathbf{d})$ and a point p in S.
Output: A spanning tree T of S rooted at p.
1. **if** $|S| = 1$
2. **then** return the tree T consisting of the single node p;
3. **else** compute the partition $S_1, S_2, \ldots, S_k$ of S and the sequence $p_1, p_2, \ldots, p_k$ of points in S, as given by Lemma 5;
4. **for** $i = 1$ **to** k
5. **do** $T_i =$ LowAverStrTree$(S_i, \mathbf{d}, p_i)$;
6. let T be the union of $T_1, T_2, \ldots, T_k$ and the edges (p, p_2), (p, p_3), $\ldots$, (p, p_k);
7. return T

In the rest of this section, we will prove that the average stretch factor of the spanning tree T that is returned by this algorithm is bounded by a constant. The following lemma generalizes Lemma 2:

Lemma 6 *Let T be the spanning tree of S that is returned by algorithm* LowAverStrTree$(S, \mathbf{d}, p)$. *For any point q in S, we have*

$$\mathbf{d}_T(p, q) \leq 3 \cdot rad_S(p).$$

Proof. Let $r = rad_S(p)$. It follows from the second claim in Lemma 5 that each of the edges $(p, p_2), (p, p_3), \ldots, (p, p_k)$ in T has length at most $2r/3$. A straightforward induction proof shows that

$$\mathbf{d}_T(p, q) \leq \sum_{j=0}^{\infty} (2/3)^j \, r = 3r.$$

∎

Lemma 7 *Assume that $|S| \geq 2$ and let T be the spanning tree of S that is returned by algorithm* LowAverStrTree$(S, \mathbf{d}, p)$. *Then*

$$ASF(T) \leq 6\gamma,$$

where γ is the constant in Lemma 5.

Proof. We will prove by induction on the size of S that

$$\sum_{\{x,y\} \in \mathcal{P}_2(S)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} \leq 6\gamma \binom{n}{2}.$$

If S contains only one element, then T is the tree consisting of one single node. In this case, the claim holds.

Assume that S contains at least two elements. The tree T is the union of the edges $(p,p_2),(p,p_3),\ldots,(p,p_k)$ and the recursively computed trees $T_1,T_2,\ldots,T_k$. Let $D = D(S_1,\ldots,S_k)$. We have

$$\sum_{\{x,y\} \in \mathcal{P}_2(S)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} = \sum_{i=1}^{k} \sum_{\{x,y\} \in \mathcal{P}_2(S_i)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} + \sum_{\{x,y\} \in D} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)}.$$

Let i be an integer with $1 \leq i \leq k$. If both x and y are in the same subset S_i, then $\mathbf{d}_T(x,y) = \mathbf{d}_{T_i}(x,y)$. Thus, by induction, we have

$$\sum_{\{x,y\} \in \mathcal{P}_2(S_i)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} = \sum_{\{x,y\} \in \mathcal{P}_2(S_i)} \frac{\mathbf{d}_{T_i}(x,y)}{\mathbf{d}(x,y)} \leq 6\gamma \binom{|S_i|}{2}.$$

For any pair $\{x,y\}$ in D, we have, by Lemma 6,

$$\mathbf{d}_T(x,y) = \mathbf{d}_T(x,p) + \mathbf{d}_T(p,y) \leq 6 \cdot rad_S(p).$$

It follows that

$$\sum_{\{x,y\} \in D} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} \leq 6 \cdot rad_S(p) \sum_{\{x,y\} \in D} \frac{1}{\mathbf{d}(x,y)} \leq 6\gamma|D|,$$

where the last inequality follows from the third claim in Lemma 5. Thus, we have

$$\sum_{\{x,y\} \in \mathcal{P}_2(S)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} \leq 6\gamma \left(\sum_{i=1}^{k} \binom{|S_i|}{2} + |D| \right).$$

Since

$$\sum_{i=1}^{k} \binom{|S_i|}{2} + |D| = \binom{n}{2},$$

it follows that

$$\sum_{\{x,y\} \in \mathcal{P}_2(S)} \frac{\mathbf{d}_T(x,y)}{\mathbf{d}(x,y)} \leq 6\gamma \binom{n}{2}.$$

∎

5 The proof of Lemma 5

Let p be an arbitrary element of S and define $r = rad_S(p)$. We have to show that there exists a partition $S_1, S_2, \ldots, S_k$ of S, for some $k \geq 2$, and a sequence $p_1, p_2, \ldots, p_k$ of points in S, such that $p_1 = p$, $p_i \in S_i$ and $rad_{S_i}(p_i) \leq 2r/3$ for all i with $1 \leq i \leq k$, and

$$\sum_{\{x,y\} \in D} \frac{1}{\mathbf{d}(x,y)} \leq \frac{\gamma}{r} |D|,$$

where γ is a constant and $D = D(S_1, \ldots, D_k)$ is the set of all unordered pairs $\{x, y\}$ in $\mathcal{P}_2(S)$ that are in different subsets of the partition.

We will construct this partition incrementally. During the construction, we maintain the following invariant:

Invariant: We have a partition of S into subsets $S_1, S_2, \ldots, S_{k-1}, R$, and a sequence $p_1, p_2, \ldots, p_{k-1}$ of points in S, such that $p_1 = p$ and for all i with $1 \leq i \leq k-1$, the following hold: $p_i \in S_i$, $rad_{S_i}(p_i) \leq 2r/3$, and

$$\sum_{x \in S_i} \sum_{y \in S_{i+1} \cup \ldots \cup S_{k-1} \cup R} \frac{1}{\mathbf{d}(x,y)} \leq \frac{3\beta}{2r} |S_i| \cdot |S_{i+1} \cup \ldots \cup S_{k-1} \cup R|,$$

where $\beta > 2$ is the constant in Lemma 1. (Thus, the constant γ will be equal to $3\beta/2$.)

Initialization: We start the construction by setting $k = 1$ and $R = S$. Then, the invariant holds.

One iteration of the construction: Assume that the invariant holds. If $R = \emptyset$, then the partition $S_1, S_2, \ldots, S_{k-1}$ of S and the sequence $p_1, p_2, \ldots, p_{k-1}$ of points prove Lemma 5.

Assume that $R \neq \emptyset$. If $k = 1$, then let $p_1 = p$. Otherwise, let p_k be an arbitrary element of R. We will show how to partition R into two subsets S_k and R' such that

$$rad_{S_k}(p_k) \leq 2r/3 \tag{4}$$

and

$$\sum_{x \in S_k} \sum_{y \in R'} \frac{1}{\mathbf{d}(x,y)} \leq \frac{3\beta}{2r} |S_k| \cdot |R'|. \tag{5}$$

Then, by setting $k = k + 1$ and $R = R'$, the invariant still holds.

First assume that all elements of R are within distance $2r/3$ of p_k. Then we define $S_k = R$ and $R' = \emptyset$. In this case, (4) and (5) obviously hold.

Thus, we may assume that not all points of R are within distance $2r/3$ of p_k. Now assume that $\{x \in R : 2r/3 < \mathbf{d}(p_k, x) \leq r\} = \emptyset$. Then we define $S_k = \{x \in R : \mathbf{d}(p_k, x) \leq 2r/3\}$ and $R' = R \setminus S_k$. It is clear that (4) holds. If $x \in S_k$ and $y \in R'$, then

$$\mathbf{d}(p_k, y) \leq \mathbf{d}(p_k, x) + \mathbf{d}(x, y),$$

which implies that $\mathbf{d}(x, y) \geq r/3$. Thus

$$\sum_{x \in S_k} \sum_{y \in R'} \frac{1}{\mathbf{d}(x, y)} \leq \frac{3}{r} |S_k| \cdot |R'|.$$

Since $\beta > 2$, it follows that (5) holds.

It remains to consider the case when $\{x \in R : 2r/3 < \mathbf{d}(p_k, x) \leq r\} \neq \emptyset$. Let $R_1 = \{x \in R : \mathbf{d}(p_k, x) \leq r\}$ and $R_2 = R \setminus R_1$. For each element x in R_1, let $r_x = \mathbf{d}(p_k, x)$. Observe that $r_{p_k} = 0$. Let $r' = \max\{r_x : x \in R_1\}$. Then $2r/3 \leq r' \leq r$. Consider the sequence of real numbers r_x, where x ranges over all elements of R_1. By Lemma 1, there exists a real number z such that $r'/3 \leq z \leq 2r'/3$ and

$$\sum_{x \in R_1, r_x \leq z} \sum_{y \in R_1, r_y > z} \frac{1}{r_y - r_x} \leq \frac{\beta}{r'} m(|R_1| - m) \leq \frac{3\beta}{2r} m(|R_1| - m),$$

where $m = |\{x \in R_1 : r_x \leq z\}|$.

We define $S_k = \{x \in R_1 : r_x \leq z\}$ and $R' = (R_1 \setminus S_k) \cup R_2$. Observe that $rad_{S_k}(p_k) \leq z \leq 2r'/3 \leq 2r/3$; thus, (4) holds. We next observe that

$$\begin{aligned}
\sum_{x \in S_k} \sum_{y \in R'} \frac{1}{\mathbf{d}(x, y)} &= \sum_{x \in S_k} \sum_{y \in R_1 \setminus S_k} \frac{1}{\mathbf{d}(x, y)} + \sum_{x \in S_k} \sum_{y \in R_2} \frac{1}{\mathbf{d}(x, y)} \\
&\leq \frac{3\beta}{2r} |S_k|(|R_1| - |S_k|) + \frac{3}{r} |S_k| \cdot |R_2| \\
&\leq \frac{3\beta}{2r} |S_k| \cdot |R'|.
\end{aligned}$$

Therefore, (5) holds. This concludes the description of one iteration of the construction.

Since in each iteration, the size of the set R gets smaller, the construction terminates. Thus, we have proved Lemma 5 and, therefore, Theorem 1 as well.

By Lemma 7, the constant α in Theorem 1 is equal to 6γ. As we have seen in the invariant, the constant γ is equal to $3\beta/2$, where β is the constant in Lemma 1. In Section 3, we have seen that Lemma 1 holds with $\beta = 4608$. Thus, we have $\alpha = 9\beta = 41,472$.

6 Concluding remarks

The minimum spanning tree (MST) may have an unbounded average stretch factor. Take $n/3$ points uniformly spaced around the unit-circle. Take two neighboring points p and q, and move them apart by a very small amount, so that their distance is a bit larger than all other distances between neighboring points. Now put $n/3$ points very close to p, and $n/3$ points very close to q. The MST of the n points is the union of (i) the unit-circle minus the gap pq, (ii) the MST of the $n/3$ points close to p, and (iii) the MST of the $n/3$ points close to q. The average stretch factor is $\Omega(n)$.

Does Theorem 1 hold for spanning *paths*? The answer is "no": Let S be the vertex set of a $\sqrt{n} \times \sqrt{n}$ grid in the plane, where each grid cell has sides of length one. Let $P = (p_1, p_2, \ldots, p_n)$ be an arbitrary spanning path of S. Let $A = \{p_1, \ldots, p_{n/3}\}$ and $B = \{p_{1+2n/3}, \ldots, p_n\}$. If $x \in A$ and $y \in B$, then $\mathbf{d}_P(x, y) \geq n/3$ and $\mathbf{d}(x, y) \leq \sqrt{2n}$. Thus,

$$ASF(P) \geq \frac{1}{\binom{n}{2}} \sum_{x \in A} \sum_{y \in B} \frac{\mathbf{d}_P(x, y)}{\mathbf{d}(x, y)} \geq \frac{1}{\binom{n}{2}} |A| \cdot |B| \frac{n/3}{\sqrt{2n}} = \Omega(\sqrt{n}).$$

Observe that we can modify algorithm EuclLowAverStrTree(S) so that it returns a spanning path of S. The analysis in Section 2, however, cannot be applied to this case, because Lemma 2 does not hold.

Acknowledgements

The author thanks all members of the Computational Geometry Lab at Carleton University for helpful discussions.

References

[1] I. Abraham, Y. Bartal, and O. Neiman. Embedding metrics into ultrametrics and graphs into spanning trees with constant average distortion. In *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, pages 502–511, 2007.

[2] N. Alon, R. M. Karp, D. Peleg, and D. West. A graph-theoretic game and its application to the k-server problem. *SIAM Journal on Computing*, 24:78–100, 1995.

[3] B. Aronov, M. de Berg, O. Cheong, J. Gudmundsson, H. Haverkort, M. Smid, and A. Vigneron. Sparse geometric graphs with small dilation. *Computational Geometry: Theory and Applications*, 40:207–219, 2008.

[4] P. B. Callahan and S. R. Kosaraju. A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields. *Journal of the ACM*, 42:67–90, 1995.

[5] G. Das and P. J. Heffernan. Constructing degree-3 spanners with other sparseness properties. *International Journal of Foundations of Computer Science*, 7:121–135, 1996.

[6] M. Elkin, Y. Emek, D. A. Spielman, and S.-H. Teng. Lower-stretch spanning trees. *SIAM Journal on Computing*, 38:608–628, 2008.

[7] D. Eppstein. Spanning trees and spanners. In J.-R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*, pages 425–461. Elsevier Science, Amsterdam, 2000.

[8] G. Narasimhan and M. Smid. *Geometric Spanner Networks*. Cambridge University Press, Cambridge, UK, 2007.