A Constant-Factor Approximation for the k-MST Problem in the Plane

Avrim Blum* Prasad Chalasani† Santosh Vempala‡

Abstract

We present an algorithm that gives a constant factor approximation for the following problem. Given a set of n points in the plane with a Euclidean distance metric and an integer $k \leq n$, find the tree of least weight that spans k points. If desired, one may also specify in the problem a "root vertex" that must be in the tree. Our result improves on the previous best bound of $O(\log k)$ of Garg and Hochbaum [5], which in turn improved a previous $O(k^{1/4})$ bound of Ravi et al [9].

1 Introduction

The k-MST problem [9] is the following. You are given a graph on n points and an integer $k \leq n$ and your goal is to find the tree of least weight that spans k points. In this paper, we consider the case that the n points are on the plane and distances are given by the Euclidean metric. Our main result is a constant factor approximation for this case, improving on the previous best bound of $O(\log k)$ by Garg and Hochbaum [5], which in turn improved on an $O(k^{1/4})$ bound of Ravi et al. [9]. For the case of general graphs, Ravi et al. [9] also had a factor $O(\sqrt{k})$ approximation, which has since been improved to $O(\log^2 k)$ [1].

*Carnegie Mellon University. Supported in part by NSF National Young Investigator grant CCR-9357793 and a Sloan Foundation Research Fellowship. avrim@cs.cmu.edu.

†Los Alamos National Laboratory. chal@lanl.gov

‡Carnegie Mellon University. Supported in part by NSF National Young Investigator grant CCR-9357793. svempala@cs.cmu.edu.

STOC '95, Las Vegas, Nevada, USA

The problem appears to have been first studied by Fischetti *et al* [4], although not from an approximation standpoint. The k-MST problem models a variety of natural situations. For example, suppose you are a salesman with a car full of k widgets, and you can sell one widget per city. You have a map of all the n cities in the U.S., but you do not want to visit every city, just enough to sell your widgets. A solution (or approximation) to the k-MST problem immediately yields an approximation to the question: what route should I take (most importantly, what cities should I visit) in order to travel as short a distance as possible.

Our algorithm is a fairly simple dynamic programming algorithm inspired by the approach of Garg and Hochbaum [5]. Our analysis involves studying the relationship between the MST for a set of points and the optimal tree of a restricted form that we call a "division tree".

2 Preliminaries and definitions

We will assume for simplicity for any set of points under consideration (in particular, for the set of n points we are given) that no two lie on the same horizontal or vertical line.

We say that a spanning tree T for a set of points P is a *Division Tree (DT)* if T satisfies the following recursive property:

> There exists some point r (the "root") such that either the vertical or the horizontal line through r splits T into two division trees. More precisely, we require both that (A) this line does not intersect any edges of T, and (B) the trees T_1 and T_2 induced by the points on either side of the line *including* r should be division trees. For the base case, if $|P| = 2$ then the single edge is a division tree.

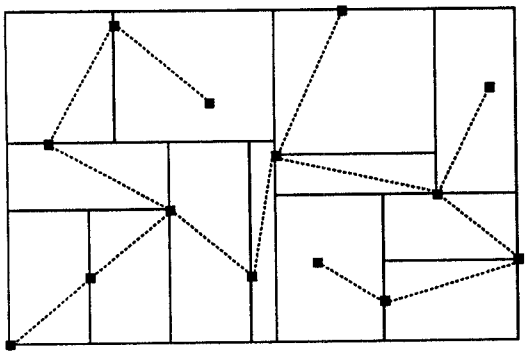

Figure 1: A division tree.

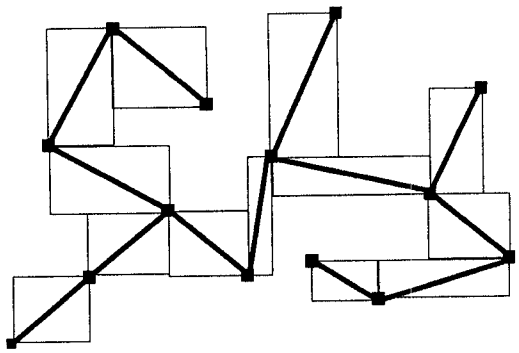

Figure 2: The final bounding boxes in the above division tree.

The *bounding box* of a set of points in the plane is the minimum enclosing axis-parallel rectangle for the set of points.

An equivalent definition of a Division Tree of a set of points P is any tree which can be constructed by the following process. Start with the bounding box B of P. Divide B with a vertical or horizontal line segment that passes through a point of P and is not one of the bounding segments of B. Construct bounding boxes for points on both sides (considering the point on the new line segment as being on both sides) and recurse on them, dividing each box using a vertical or horizontal segment and so on. Continue until each box has exactly two points (which will be at opposite corners of the box). Finally, connect the two points in each box. It is not hard to see that the set of edges added by this procedure will form a spanning tree of P and that this is equivalent to the earlier definition. We will often identify the division tree with the $|P| - 1$ boxes created. Figure 1 shows an example of a division tree, and Figure 2 shows the final bounding boxes.

The notion of partitioning a collection of points with dividing lines has appeared elsewhere in the literature. For instance Karp [7] uses a structure resembling a division tree to approximate the optimal traveling salesman tour for points randomly scattered in a rectangular region of the plane. Gonzalez and Zheng [6] use "guillotine partitions" (which are similar to our dividing lines) for a different approximation problem.

3 Algorithm

Our algorithm for the k-MST problem uses a dynamic programming procedure to find the subset of k points with a division tree of minimum weight.

The algorithm is most easily viewed in a recursive "memoizing" form. It returns both the desired set of k points and the *cost* of the associated division tree. The algorithm takes as input a set of points P, an integer k, and also up to four additional constraints. For each of the 4 sides of the bounding box of P the algorithm may be told that the point on that bounding side is "required" and must be in any set of k points the algorithm produces. At the outer loop there are no required points. Given these inputs, the algorithm considers each vertical and horizontal line that passes through some point in P (actually, lines that coincide with an edge of the bounding box need not be considered). For a given such line—let p be the point in P that the line passes through—the algorithm constructs the bounding boxes B_1 and B_2 of the points in P on the two sides, considering p to be on both sides. It then calls itself recursively $k-1$ times for each of the two boxes B_i: in each call passing down the set of points in B_i, a new integer $k' \in [2,k]$, and the set of required points it was originally given (only considering those that lie in the box B_i) *including* the new point p. Once the algorithm receives its $k-1$ answers from each side, it simply compares to find the pair $\langle k', k-k'+1 \rangle$ whose costs sum to the least amount (the reason for the "+1" is that point p lies on both sides). In the base case, $k = 2$, the algorithm just returns

the cost of the single edge.

Because there are at most n^4 different bounding boxes, k different possibilities for the desired number of points, and 16 different settings for the "required points", the memoized procedure (or equivalently, dynamic program) will run in polynomial time.

It is not too hard to see from the construction and the definition of division tree that this algorithm finds the set of k points with the lightest division tree. What remains to be shown (and what the sections below consist of a proof of) is that for any set of points, the division tree of minimum weight is only a constant factor more costly than the minimum spanning tree.

4 Lower-Bounding Technique

The main result of this paper is the following geometric fact about points in the plane.

Theorem 1 *Let P be a set of points in the Euclidean plane. Then there is a constant c so that $MDT(P) \leq c \cdot MST(P)$, where MDT and MST stand for minimum division tree and minimum spanning tree respectively.*

In our discussion, the length of a rectangle refers to its longer side and the width to its shorter side. In particular, for box B, $length(B)$ and $width(B)$ are the lengths of the longer and shorter sides of B respectively. An axis-parallel rectangle is called *horizontal (vertical)* if its length is horizontal (vertical). We define the *centre* of a rectangle to be the intersection of its diagonals.

Definition 1 *An* r-Fat box *($r \leq 1$) is a rectangle with the ratio of its smaller side to larger side at least r. An* r-Thin box *is a rectangle which is not r-fat.*

In section 5 we will fix a specific value of r and abbreviate r-fat and r-thin to just fat and thin respectively. To derive lower bounds on the minimum spanning tree (MST) of points on the plane, we make use of a theorem due to Das et al [3, 2].

Theorem 2 (Existence of Spanners) *Let P be a set of points in the plane. Then for every $\epsilon > 0$ there exists a Steiner graph T on P (by Steiner graph we mean a graph on a superset of P) that is light and distance-preserving in the following sense:*

1. *The sum of the weights of edges in T is at most $g(\epsilon) \cdot MST(P)$ and $g(\epsilon)$ is a quantity that depends only on ϵ.*
2. *The distance* in T *(i.e. using only edges in T) between any two points $u, v \in P$ is at most $(1+\epsilon)d(u,v)$ where $d(,)$ is Euclidean distance.*

Lemma 1 (Fat-Box Lower Bound) *Let F be a set of r-fat boxes with non-overlapping interiors, with each box $f \in F$ specifying two points u_f and v_f on its boundary such that the line joining them intersects the centre of the box. Let $P(F) = \bigcup_{f \in F}\{u_f, v_f\}$ denote the set of these boundary points. Then $MST(P(F)) \geq c \cdot \Sigma_{f \in F}(length(f))$ where c is a constant that depends only on r.*

Proof. Use Theorem 2 to build a light distance-preserving graph on $P(F)$ with ϵ chosen so that for each $f \in F$, at least half of the shortest path in T between u_f and v_f lies inside f (for instance $\epsilon = \frac{\frac{1}{2}+\sqrt{\frac{1}{4}+r^2}-\sqrt{1+r^2}}{\sqrt{1+r^2}}$). In other words since f is r-fat, any path between u_f and v_f that has less than half its length inside f must be longer than $(1+\epsilon)d(u_f, v_f)$. This implies that the graph T constructed by the theorem has total edge weight at least half the sum of the lengths of all boxes in F. Therefore the statement of the theorem is true with $c = 1/2g(\epsilon)$. ∎

An implication of this lemma for proving Theorem 1 is that if we are lucky and there exists a division tree for P such that all the $|P|-1$ boxes created are fat, then we are done. This is because the cost of the division tree is at most $\sqrt{2}$ times the sum of the lengths of all the boxes.

The above lemma can be generalized in the following way: remove the restriction that the fat boxes be non-overlapping, and instead each fat box is required to have a "large" empty region such that these regions do not overlap. Then by choosing ϵ small enough (a constant depending on the size of the empty region) in the construction of a spanner for $P(F)$ we can ensure that the path

from u to v in a box f intersects at least half the empty region. The lemma also remains true if we relax the requirement that each box has two points on the ends of a diagonal to the requirement that there should be a point on each bounding line of each fat box. The following refinement of the fat-box lower bound incorporates these generalizations:

Lemma 2 (Empty-Regions Lower Bound) *Let F be a set of r-fat boxes with the following properties (see Figure 3):*

1. *Each box $f \in F$ has a point on every bounding line. Let $P(f)$ denote this set of points.*
2. *Each box $f \in F$ contains "large" empty regions, (where by empty we mean that the interior has no points): for constants $a, b > 0$, either*
   (a) *f has one or more empty strips (rectangles extending fully along its width) such that the sum of the sizes of the strips in f along the length of f is at least $a \cdot length(f)$ (in other words, the strips together contain a constant fraction 'a' of the area of f), OR*
   (b) *f has an axis-parallel r-fat empty rectangle of length at least $b \cdot length(f)$ whose centre is on a line joining two of the boundary points of f.*
3. *The empty regions (strips or rectangles) of two different boxes are non-overlapping, i.e. for any two boxes $f_1, f_2 \in F$, and any two empty regions $c_1 \in f_1$ and $c_2 \in f_2$, c_1 and c_2 are non-overlapping.*

Let $P(F) = \bigcup_{f \in F} P(f)$. Then $MST(P(F)) \geq c \sum_{f \in F}(length(f))$ where $c > 0$ is a constant that depends only on r, a and b.

Proof. For the proof we define an auxiliary collection X of fat boxes and two relevant points u_g and v_g for each $g \in X$. Let F be a collection of fat boxes satisfying the requirements of the lemma and let $f \in F$. Assume f is horizontal. Let the points on the left, top, right and bottom boundaries be p_l, p_t, p_r, p_b respectively (these are not necessarily distinct). For each f we look at the following cases:

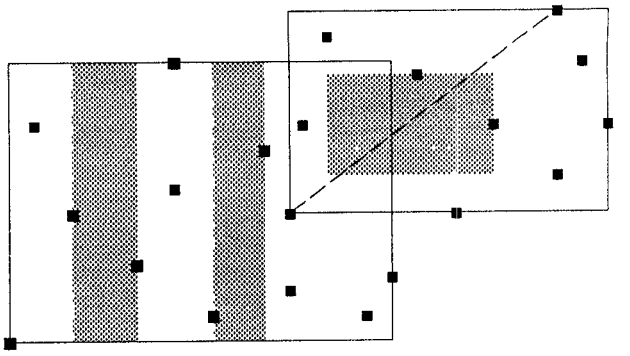

Figure 3: Illustrating the Empty-region lower bound. Empty regions are shown shaded.

1. If f satisfies condition 2b then we put f in X. Define u_f and v_f to be the two points on the boundary of f from condition 2b.
2. Otherwise, if the bounding box of $\{p_l, p_r\}$ is $\frac{r}{3}$-fat, let f_1 be this bounding box. We place f_1 in X with $u_{f_1} = p_l$ and $v_{f_1} = p_r$.
3. Otherwise, if p_l is in the middle third of the left boundary of f or if p_r is in the middle third of the right boundary of f, then we can draw an axis-parallel $\frac{r}{3}$-fat box f_1 inside f of the same length so that the line joining p_l and p_r intersects the centre of the box. Then add f_1 to X with $u_{f_1} = p_l$ and $v_{f_1} = p_r$.
4. Otherwise p_l and p_r are both close to the top or bottom boundary of f. Assume they are close to the bottom boundary of f. Consider the bounding boxes of $\{p_l, p_t\}$ and $\{p_t, p_r\}$. One of them (call it f_1) must contain at least half the width of the empty strips and is therefore $min(a/2, 2r/3)$-fat. Then $f_1 \in X$ with u_{f_1}, v_{f_1} set accordingly.

This collection of fat boxes X has the property that each box f of length l either satisfies condition 2b or is $min(a/2, r/3)$-fat and has two points on its boundary so that the line joining them passes through the centre of the box and has empty strips totaling to at least $a \cdot l/2$. In each box f of length l draw two lines parallel to the line joining u_f, v_f at a distance $min(a \cdot l/8, b \cdot l/4)$ on either side. Use Theorem 2 with ϵ small enough so that the path from u to v in each box does not cross these

parallel lines. This can be done with ϵ as a function of only a, b, r. This ensures that the path intersects at least a constant fraction of the empty region in each box in X and completes the proof. ∎

5 Analysis

The algorithm presented earlier finds the subset of k points (from the given set of n points) that has the minimum weight division tree among all k point subsets. In the rest of the paper we show that for any set of points there exists a division tree whose weight is at most a constant times the minimum spanning tree for the set.

Let P be a set of points in the plane. The final set of boxes created by finding a division tree for P, i.e, boxes with no further subdivision and two points per box, will be called *leaves*. Let this set of boxes be denoted by L. As noted above, it is easy to see that $DT(P) \leq \sqrt{2}\sum_{l \in L}(length(l))$; so if all the leaves obtained are fat, using the fat-box lower bound we have $DT(P) \leq O(1) \cdot MST(P)$. However this is not the case in general.

5.1 Constructing a light division tree

To prove the mathematical statement that there exists a light division tree, we present a construction whose basic approach will be to create r-fat boxes if possible, where we use $r = 1/10$. Here is the precise construction: We start with the bounding box of the given set of points. The general step is that we arbitrarily pick a box B which has more than two points in it (at the first step it is the original bounding box) and divide B by adding one or more lines. Assume that B has its length horizontal. When dividing B vertically through a point u we denote the left and right bounding boxes by $B_l(u)$ and $B_r(u)$ respectively. If B is thin, then simply add the vertical line through the point horizontally closest to the center of the longer side and collapse both sides to bounding boxes. If B is fat, then if possible divide it with one or more vertical lines so that on forming bounding boxes all the boxes obtained are fat. If, however, this is not possible, then there are three cases to be considered:

1. Let the point horizontally closest to the center of the longer side be u. Consider the vertical line through u. If the horizontal side of $B_l(u)$ or $B_r(u)$ is smaller than $length(B)/4$, then add the two vertical lines which mark out the empty region of horizontal side at least $length(B)/2$ in the middle of the box. This is possible unless the empty region extends all the way to the right end of B, i.e. $B_r(u)$ is a leaf. In this case just add the one line considered. (Note that $B_l(u)$ and $B_r(u)$ may be thin, horizontally or vertically)

2. Otherwise, if B is $5r$-fat, add the vertical line through u, the point horizontally closest to the center.

3. Otherwise, again consider the vertical line through u. It must induce a horizontal thin box on one side. Assume $B_l(u)$ is horizontal thin. Then start at the left end of B and locate the first point x from the left which has the property that the $B_l(x)$ is horizontal thin. Add vertical lines through x and the point y which is just to the left of x. If x is the second point from the left then there is only one line to be added.

The above division rules are illustrated in Figure 4. We make a few useful observations about this construction.

Claim 1 *In case 2 of the construction, assume w.l.o.g. that $B_l(u)$ is horizontal thin (at least one of $B_l(u)$ or $B_r(u)$ must be) and let d be its length. Then there is an r-fat empty rectangle above or below $B_l(u)$ of length at least $3d/5$ so that its centre lies on the line joining two of the boundary points of B. (See Figure 5.)*

Proof. $B_l(u)$ is horizontal thin. Let v be a point on the left boundary of B (also on the left boundary of $B_l(u)$) and assume without loss of generality that v is closer to the bottom of B than to the top. Let w be a point on the top boundary of B. Let p_1 be the point where the line joining v and w intersects the vertical line through u and let p_2 be the point where this line intersects the top boundary of $B_l(u)$. Notice that the slope of the line vw is at

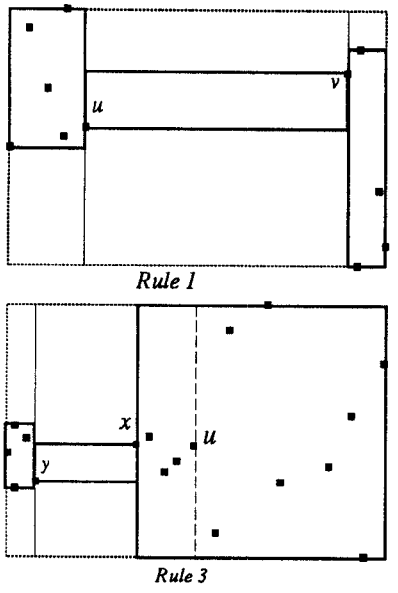

Figure 4: Division rules.

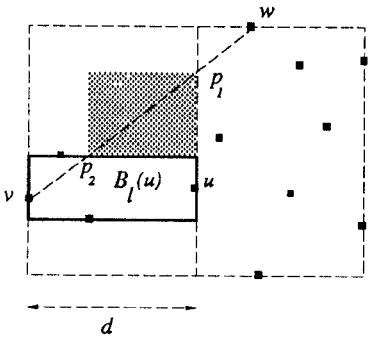

Figure 5: Illustrating Claim 1.

least $\frac{5r}{2}$. Thus, the vertical distance of p_1 from the top boundary of $B_l(u)$ is at least $\frac{5r}{2} \cdot d - r \cdot d = \frac{3r}{2} \cdot d$. Also the horizontal distance from p_2 to the vertical line through u is at least $d - rd\frac{2}{5r} = 3d/5$. We can draw a rectangle of these dimensions (namely, the bounding box of $\{p_1, p_2\}$) whose centre is on the line joining v and w. Since $d \geq length(B)/4$ the claim follows. ■

Claim 2 *In case 3 of the construction, the points x and y will be such that their bounding box has its length horizontal.*

Proof. Let x and y be the two points chosen by case 3 of the construction. $B_l(y)$ is fat or vertical thin. If it is vertical thin, then since $B_l(x)$ is horizontal, the bounding box of $\{x, y\}$ is horizontal. If $B_l(y)$ is fat then $B_r(y)$ must be horizontal thin, as is $B_l(x)$ by definition of x. Denoting the horizontal side of the bounding box of $\{x, y\}$ as *horiz* and the vertical side as *vert*, we have

$$\begin{aligned} length(B) &= length(B_l(x)) + length(B_r(y)) \\ &\quad -horiz, \\ width(B) &\leq width(B_l(x)) + width(B_r(y)) \\ &\quad -vert. \end{aligned}$$

Since B is fat but $B_l(x)$ and $B_r(y)$ are horizontal thin, this implies that the bounding box of $\{x, y\}$ is horizontal. ■

Claim 3 *Suppose that in case 3, $B_l(y)$ is vertical thin. Let the horizontal distance between x and y be d. Then $width(B_l(y)) \leq d\frac{r^2}{1-r^2}$, which is (much) less than $dr/2$ for our chosen value of $r = 1/10$.*

Proof. Follows from the fact that $B_l(x)$ is horizontal thin. ■

5.2 Bounding the weight

Let P be a set of k points on the plane. Construct a division tree for P using the procedure of the previous section. Let this be T. The weight of T can be written as the sum of the weight in fat leaves and the weight in thin leaves. From the fat-box lower bound we know that the weight of T

in fat leaves is at most a constant times MST(P). The main lemma of the proof puts a bound on the weight of T in thin leaves.

Lemma 3 (The main Lemma) *The weight of T in thin leaves is less than $c \cdot MST(P)$ for a fixed constant c.*

In order to prove this lemma, we need to consider two different kinds of thin boxes. Let us define the *parent* of a box B to be the box whose subdivision resulted in the immediate creation of B. Then a thin box can be classified into one of two types depending on whether its parent is fat or thin. F-thin boxes will be those with fat parents and T-thin boxes will be those with thin parents. First we account for the weight in T-thin leaves. The weight of the edge in a T-thin leaf is charged to its parent. If this thin parent itself has a thin parent (i.e., it is a T-type box) then it will in turn pass all its charge to its parent, and so on. This way the weight of T-thin leaves will reach F-thin boxes which are not leaves (if the original box—the bounding box of P—is thin, we will label it as F-thin).

We define the charge on a box to be the total charge it receives from its descendents. Also, for convenience, we say that a thin leaf has charge equal to the weight of its edge. The next claim upperbounds the total charge that can accumulate on a thin box by this process.

Claim 4 *The total weight of T-thin leaves charged to a thin box is less than twice the length of the thin box.*

Proof. Let t be a thin box. Assume without loss of generality that t is horizontal thin. Explore the subdivision of t stopping whenever a fat box is reached. We prove the claim by induction on the number of points in t. For 2 points (t is a leaf) the claim is trivially true (charge = length of edge joining the two points). Consider now the general case and let t_1, t_2 be the boxes obtained on dividing t with the next line segment of the division tree. If one of t_1, t_2, say t_1, is fat then $charge(t) \leq charge(t_2)$ and we are done by the induction hypothesis as a fat box does not send up any charge. If both t_1 and t_2 are horizontal thin, then by the hypothesis $charge(t) \leq charge(t_1) + charge(t_2) \leq 2(length(t_1) + length(t_2)) \leq 2 \cdot length(t)$. Lastly, If one of them, say t_1 is vertical thin, then the next subdivision of t_2 will produce a leaf l and possibly another box t_3. Then $length(t_1) \leq r \cdot length(t)$ and $length(t_3) \leq r \cdot length(t)$, giving us $charge(t) \leq charge(t_1) + charge(t_3) + (\sqrt{1+r^2})length(t) \leq (4r + \sqrt{1+r^2})length(t)$, which is less than $2 \cdot length(t)$ for $r \leq 0.24$. ■

By the above claim, all that remains is to bound the total length of the F-thin boxes by some constant times the cost of the MST. Let us now consider the ways in which such boxes can be created.[1] First, consider F-thin boxes created in case 2 of the construction from a $5r$-fat parent. By Claim 1 the F-thin boxes created in case 2 have total length at most a constant times the cost of the MST since their parent boxes satisfy condition 2b of the Empty-Regions lower bound. Now consider cases 1 and 3. In each of these cases, a large empty strip is created (between u and v in case 1 and between x and y in case 3) whose width is at least half the total length of any thin boxes created. Thus, all we need to show is that the sum of the widths of the empty strips is at most a constant times the MST of P, and this will bound the total length of the F-thin boxes.

To do this we use a charging procedure as follows. The idea is that we are going to charge the widths of the empty strips to fat boxes, which will in turn pass their charge down to their children, until all the charge resides on boxes that can be used for Lemma 2 (the Empty-Regions Lower Bound). We use X to denote the set of these useful boxes. For fat boxes *not* in X we maintain the invariant that each is charged at most its length + width, and for fat boxes in X the charge will only be a constant factor larger. Thus, Lemma 2 will imply that the total charge is only a constant factor greater than the cost of the MST.

Consider a fat box f with perhaps some charge. We now examine the cases for division of f in our construction. Let $a = 1/10$.

[1] One minor point: the (initial) bounding box of P, if it is thin, can be accounted for separately since its length is clearly at most the cost of the MST.

0. If we divide f into two fat boxes f_1 and f_2, we transfer the charge on f (if any) to f_1 and f_2 in proportion to their perimeters. Notice that the length+width of f is at most the length+width of f_1 plus the length+width of f_2.

1. If we divide f using rule 1 of the construction, then f has a large empty strip (width at least half the length of f). So we place f into X and we charge the width of the empty strip to f as well.

2. If we use rule 2 to divide f (f was $5r$-fat), we again just put f into X with its given charge.

3. If we used rule 3 to divide f then we continue subdividing with vertical lines every horizontal fat box created (except those $5r$-fat that fall into rule 2 of the construction) until every box within f is either a vertical (fat or thin) box, a leaf (inducing a vertical empty strip), or a $5r$-fat box falling under rule 2. There are now 3 subcases.

   (a) If the sum of the widths of the empty strips is at least $a/2 \cdot length(f)$, we place f into X. We charge to f the widths of the empty strips (totaling to at most the length of f).

   (b) If the sum of the lengths of the $5r$-fat boxes in f is at least $a/2 \cdot length(f)$ we place these $5r$-fat boxes into X, distributing among them the charge on f and the sum of the widths of the empty strips.

   (c) Otherwise (the sum of the widths of strips in f and the lengths of the $5r$-fat boxes is less than $a \cdot length(f)$) let S be the set of vertical fat boxes obtained. We now charge the widths of the strips and the charge on f to the boxes in S in proportion to their perimeters.

   The point of this is that in this case the boxes in S must have a large total length+width since they are all vertical and have large total width (since at most $a \cdot length(f)$ of the length of f is used by empty strips or $5r$-fat boxes).

The last claim bounds the total charge that can reach a fat box.

Claim 5 *Fix $a = 1/10$ and $r = 1/10$. Then the total charge on a box is less than the length+width of the box.*

Proof. By induction. Assume f is a horizontal fat box.

If f is divided into two fat boxes then this charge is transferred to the two boxes and the charge per unit length remains the same or decreases. The only other time charge is passed down is in the final case (3c) above. Let $l = length(f)$. In this case $\sum_{s \in S} width(s) \geq (1 - a - a/2 - ar/2)l$, (widths of empty strips plus lengths of $5r$-fat boxes total to less than al by assumption, thin boxes adjacent to empty strips created in case 3 of the construction are of total width at most $(ar/2)l$ by Claim 3, and thin boxes created in case 1 of the construction, horizontal or vertical, are of length at most $(a/2)l$ by the construction). Also $(length(f) + width(f)) \leq l(1 + 5r)$.

Combining the last two facts we have: the total charge to boxes in S divided by the sum of the length plus widths of the boxes in S is at most $(l(1 + 5r) + al)/(2l(1 - a - a/2 - ar/2))$. Using $a = r = 1/10$ this is less than 1. ■

We now combine the above claims to prove the main lemma.

Proof of Lemma 3. As shown above, all the charge reaches boxes which are added to X. Using the Empty-Regions lower bound the (length+width) of boxes in X is a lower bound on the cost of the minimum spanning tree of P (up to a constant). This proves the main lemma :

Total weight of $DT(P)$ in thin leaves
$\leq$ $O(1) \cdot$ [total length of F-thin boxes] (by Claim 4)
$\leq$ $O(1) \cdot$ [total charge on fat boxes in X]
$\leq$ $O(1) \cdot$ [total length of fat boxes in X] (by Claim 5)
$\leq$ $O(1) \cdot MST(P)$ (by Lemma 2)

■

Theorem 1 is immediate from the main lemma.

Final comments

Although the constant resulting from our proof is quite high (and depends on the constant from Das et al.), the algorithm itself appears to be very good. We have not been able to find any examples on which it performs worse than a factor of 2.

Very recently R. Ravi has an exciting new result that appears to achieve a constant factor approximation for the k-MST problem in general graphs [8].

References

[1] B. Awerbuch, Y. Azar, A. Blum, and S. Vempala. Improved approximation guarantees for minimum-weight k-trees and prize-collecting salesmen. Technical Report CMU-CS-94-173.

[2] G. Das, P. Heffernan, and G. Narasimhan. Optimally sparse spanners in 3-dimensional Euclidean space. In *9th Annual ACM Computational Geometry*, pages 53–62, 1993.

[3] G. Das, G. Narasimhan, and J. Salowe. A new way to weigh malnourished euclidean graphs. In *Symp. on Discrete Algorithms (SODA)*, 1994.

[4] M. Fischetti, H.W. Hamacher, K. Jornsten, and F. Massioli. Weighted k-cardinality trees: complexity and polyhedral structure. *Networks*, 24(1):11–21, 1994.

[5] N. Garg and D. Hochbaum. An $O(\log k)$ approximation for the k minimum spanning tree problem in the plane. In *STOC*, 1994.

[6] T. Gonzalez and S. Zheng. Improved bounds for rectangular and guillotine partitions. *J. Symbolic Computation*, 7:591–610, 1989.

[7] R. M. Karp. Probabilistic analysis of partitioning algorithms for the traveling-salesman problem in the plane. *Math. of O.R.*, 2(3):209–224, 1977.

[8] R. Ravi. Personal communication.

[9] R. Ravi, R. Sundaram, M. Marathe, D. Rosenkrantz, and S. S. Ravi. Spanning trees short or small. In *Symp. on Discrete Algorithms (SODA)*, 1994.