SIAM J. COMPUT.
Vol. 44, No. 1, pp. 160–192

EMBEDDING METRICS INTO ULTRAMETRICS AND GRAPHS INTO SPANNING TREES WITH CONSTANT AVERAGE DISTORTION*

ITTAI ABRAHAM†, YAIR BARTAL‡, AND OFER NEIMAN§

Abstract. This paper addresses the basic question of how well a tree can approximate distances of a metric space or a graph. Given a graph, the problem of constructing a spanning tree in a graph which strongly preserves distances in the graph is a fundamental problem in network design. We present scaling distortion embeddings where the distortion scales as a function of ϵ, with the guarantee that for each ϵ simultaneously, the distortion of a fraction $1-\epsilon$ of all pairs is bounded accordingly. Quantitatively, we prove that any finite metric space embeds into an ultrametric with scaling distortion $O(\sqrt{1/\epsilon})$. For the graph setting, we prove that any weighted graph contains a spanning tree with scaling distortion $O(\sqrt{1/\epsilon})$. These bounds are tight even for embedding into arbitrary trees. These results imply that the *average distortion* of the embedding is constant and that the ℓ_2 distortion is $O(\sqrt{\log n})$. For probabilistic embedding into spanning trees we prove a scaling distortion of $\tilde{O}(\log^2(1/\epsilon))$, which implies *constant* ℓ_q-distortion for every fixed $q < \infty$.

Key words. embedding, constant average distortion, spanning tree

AMS subject classification. 30L05

DOI. 10.1137/120884390

1. Introduction.

The problem of embedding general metric spaces into tree metrics with small distortion has been central to the modern theory of finite metric spaces. Such embeddings provide an efficient representation of the complex metric structure by a very simple metric. Moreover, the special class of ultrametrics (rooted trees with equal distances to the leaves) plays a special role in such embeddings [Bar96, BLMN05]. Such an embedding provides an even more structured representation of the space which has a hierarchical structure [Bar96]. Probabilistic embedding into ultrametrics has led to algorithmic applications for a wide range of problems (see [Ind01]).

An important problem in network design is to find a tree spanning the network, represented by a graph, which provides good approximation of the metric defined with the shortest path distances in the graph. Different notions have been suggested to quantify how well distances are preserved, e.g., routing trees and communication trees [WLB+98]. The papers [AKPW95, EEST05] study the problem of constructing a spanning tree with low average stretch, i.e., low average distortion over the edges of the tree. It is natural to define our measure of quality for the embedding to be its average distortion over all pairs, or alternatively the more strict measure of its

*Received by the editors July 12, 2012; accepted for publication (in revised form) November 20, 2014; published electronically February 18, 2015. A preliminary version of this paper appeared in *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2007.
http://www.siam.org/journals/sicomp/44-1/88439.html
†VMware Research Group, 3401 Hillview Avenue, Palo Alto, CA 94304 (ittaia@gmail.com, iabraham@vmware.com). Part of this work was done while this author was at Microsoft Research Silicon Valley.
‡Computer Science School, Hebrew University, Jerusalem, Israel (yair@cs.huji.ac.il). This author was supported in part by a grant from the Israeli Science Foundation (1609/11).
§Department of Computer Science, Ben-Gurion University of the Negev, Beer Sheva 84105, Israel (neimano@cs.bgu.ac.il). This author was supported in part by ISF grant 523/12 and by the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement 303809.

ℓ_2-distortion. Such notions are very common in most practical studies of embeddings (see, for example, [HS00, HFC00, AS03, HBK+03, ST04, TC04]). We recall the definitions from [ABN06]. Given two metric spaces (X, d_X) and (Y, d_Y) an *injective* mapping $f : X \to Y$ is called an *embedding* of X into Y. An embedding is *noncontractive* if for any $u \neq v \in X$: $d_Y(f(u), f(v)) \geq d_X(u, v)$. For a noncontractive embedding let the distortion of the pair $u, v \in X$ be $\mathrm{dist}_f(u,v) = \frac{d_Y(f(u),f(v))}{d_X(u,v)}$.

Definition 1.1 (ℓ_q-distortion). *For $1 \leq q \leq \infty$, define the* ℓ_q-distortion *of an embedding f as*

$$\mathrm{dist}_q(f) = \mathbb{E}[\mathrm{dist}_f(u,v)^q]^{1/q},$$

where the expectation is taken according to the uniform distribution over $\binom{X}{2}$. The classic notion of distortion *is expressed by the ℓ_∞-distortion and the* average distortion *is expressed by the ℓ_1-distortion.*

Besides $q = \infty$ and $q = 1$, the case of $q = 2$ is a natural measure. It is related to the notion of *stress* which is a standard measure in *multidimensional scaling* methods [KW78, Kru64]. See [ABN06] for more information regarding the ℓ_q distortion.

Definition 1.2 (partial/scaling embedding). *Given two metric spaces (X, d_X) and (Y, d_Y), a* partial embedding *is a pair (f, P), where f is a noncontractive embedding of X into Y, and $P \subseteq \binom{X}{2}$. The distortion of (f, P) is defined as* $\mathrm{dist}(f, P) = \sup_{\{u,v\} \in P} \mathrm{dist}_f(u, v)$. *For $\epsilon \in [0, 1)$, a* $(1-\epsilon)$-partial embedding *is a partial embedding such that $|P| \geq (1-\epsilon)\binom{n}{2}$.*[1] *Given two metric spaces (X, d_X) and (Y, d_Y) and a function $\alpha : [0, 1) \to \mathbb{R}^+$, we say that an embedding $f : X \to Y$ has* scaling distortion *α if for any $\epsilon \in [0, 1)$, there is some set $P(\epsilon)$ such that $(f, P(\epsilon))$ is a $(1-\epsilon)$-partial embedding with distortion at most $\alpha(\epsilon)$.*

The notion of average distortion is tightly related to that of embedding with scaling distortion [KSW04, ABC+05], as shown by the following lemma proved in [ABN06].

Lemma 1.3. *Given an n-point metric space (X, d_X) and a metric space (Y, d_Y). If there exists an embedding $f : X \to Y$ with scaling distortion α, then*

$$\mathrm{dist}_q(f) \leq \left(2 \int_{\frac{1}{2}\binom{n}{2}^{-1}}^{1} \alpha(x)^q dx \right)^{1/q}.$$

We prove the following theorems.

Theorem 1.4. *Any n-point metric space embeds into an ultrametric with scaling distortion $O(\sqrt{1/\epsilon})$. In particular, its ℓ_q-distortion is $O(1)$ for any fixed $1 \leq q < 2$, $O(\sqrt{\log n})$ for $q = 2$, and $O(n^{1-2/q})$ for any fixed $2 < q \leq \infty$.*

Theorem 1.5. *Any weighted graph of size n contains a spanning tree with scaling distortion $O(\sqrt{1/\epsilon})$. In particular, its ℓ_q-distortion is $O(1)$ for any fixed $1 \leq q < 2$, $O(\sqrt{\log n})$ for $q = 2$, and $O(n^{1-2/q})$ for any fixed $2 < q \leq \infty$.*

The tightness of our results follows from a lower bound in [ABC+05]. We show in section 5 that the bounds in Theorems 1.4 and 1.5 are tight for the n-cycle, even for embeddings into arbitrary tree metrics.

A probabilistic embedding is a distribution $\mathcal{F}$ over noncontracting embeddings, and the distortion of the pair $\{u, v\}$ is $\mathrm{dist}_{\mathcal{F}}(u, v) = \mathbb{E}_{f \in \mathcal{F}}[\frac{d_Y(f(u),f(v))}{d_X(u,v)}]$. The notion

[1] Note that the embedding is *strictly* partial only if $\epsilon \geq 1/\binom{n}{2}$.

of scaling distortion is extended to probabilistic embedding in the obvious way. We obtain an equivalent result for probabilistic embedding into spanning trees.

THEOREM 1.6. *Any weighted graph of size n probabilistically embeds into a spanning tree with scaling distortion $\tilde{O}(\log^2 1/\epsilon)$. In particular, its ℓ_q-distortion is $O(1)$ for any fixed $1 \le q < \infty$.*[2]

1.1. Related work. Embedding metrics into trees and ultrametrics was introduced in the context of probabilistic embedding in [Bar96]. Other related results on embedding into ultrametrics include work on metric Ramsey theory [BLMN05], multi-embeddings [BM03], and dimension reduction [BM02]. Embedding an arbitrary metric into a tree metric requires $\Omega(n)$ distortion in the worst case even for the metric of the n-cycle [RR98]. It is a simple fact [HPM06, BLMN05, Bar96] that any n-point metric embeds in an ultrametric with distortion $n-1$. However, the known constructions are not scaling and have average distortion linear in n.

The probabilistic embedding theorem [FRT03, Bar04] (improving earlier results of [Bar96, Bar98]) states that any n-point metric space probabilistically embeds into an ultrametric with distortion $O(\log n)$. This result has been the basis for many algorithmic applications (see [Ind01]). This theorem implies the existence of a single ultrametric with average distortion $O(\log n)$ (a constructive version was given in [Bar04]). This bound was later improved with the analysis of [ABC$^+$05], as we discuss below.

The study of partial embedding and embedding with scaling distortion was initiated by Kleinberg, Slivkins, and Wexler [KSW04] and later studied in [ABC$^+$05, ABN06]. Abraham et. al [ABC$^+$05] prove that any finite metric space probabilistically embeds in an ultrametric with scaling distortion $O(\log(1/\epsilon))$ implying constant average distortion. As mentioned above, since the distortion is bounded in expectation, this result implies the existence of a single ultrametric with constant average distortion but does not bound the ℓ_2-distortion. In [ABN06] we have studied in depth the notions of average distortion and ℓ_q-distortion and their relation to partial and scaling embeddings. Our main focus was the study of optimal scaling embeddings for embedding into L_p spaces. For embedding of metrics into ultrametrics, we mentioned that *partial* embeddings exist with distortion $O(\sqrt{1/\epsilon})$ matching the lower bound from [ABC$^+$05]. Theorem 1.4 significantly strengthens this result by providing an embedding with *scaling* distortion. That is, the bound holds for all values of $0 < \epsilon < 1$ *simultaneously* and therefore the embedding has bounded ℓ_q-distortion as given by Lemma 2.2.

It is a basic fact that the minimum spanning tree (MST) in an n-point weighted graph preserves the (shortest paths) metric associated with the graph up to a factor of $n-1$ at most. This bound is tight for the n-cycle. Alas, in general the MST does not have scaling distortion and may have linear average distortion.[3] Alon et al. [AKPW95] studied the problem of computing a spanning tree of a graph with small average stretch (over the edges of the graph). This can also be viewed as the dual of probabilistic embedding of the graph metric in spanning trees. Their work was significantly improved by Elkin et al. [EEST05], who show that any weighted graph contains a spanning tree with average stretch $O(\log^2 n \log\log n)$. Further improvements by [ABN08, AN12] gave a near optimal $\tilde{O}(\log n)$ bound. This result can also be rephrased in terms of the average distortion (but not the ℓ_2-distortion) over all

[2] Note that probabilistic embedding bounds on the ℓ_q-distortion *do not* imply an embedding into a single tree with the same bounds, with the exception of $q = 1$.

[3] For example, take the uniform metric and slightly perturb it, so that the MST is a path.

pairs. For spanning trees, this paper gives the *first* construction with *constant* average distortion.

We remark that the result of Theorem 1.6 was obtained before the improvements of [ABN08, AN12], and while it seems possible that it could be improved to $\tilde{O}(\log(1/\epsilon))$ using the new ideas of these papers, there are some technical complications, and therefore we have decided not to pursue this direction here.

Following our work, [ELR07] showed that Theorem 1.5 resolves a conjecture of [DPK82] from 1982 on cycle bases. The weight of a strictly fundamental cycle basis for a spanning tree T of a graph $G = (V, E)$ is essentially $\sum_{(u,v)\in E} d_T(u, v)$ (up to a factor of 2), and Deo, Prabhu and Krishnamoorthy conjectured that for any unweighted graph there exists a spanning tree for which this quantity is bounded by $O(n^2)$. Our result gives a stronger bound, that $\sum_{(u,v)\in E} d_T(u, v) + \sum_{(u,v)\notin E} \frac{d_T(u,v)}{d(u,v)} \le O(n^2)$.

1.2. Discussion of techniques. Theorem 1.4 uses partitioning techniques similar to those used in the context of the metric Ramsey problem [BBM06, BLMN05]. However, in our case we need to provide an argument for the existence of a partition which simultaneously satisfies multiple conditions, each for every possible value of ϵ. Theorem 1.5 builds on the technique above together with the Elkin et al. [EEST05] method to construct a spanning tree. A straightforward application of this approach loses an extra $O(\log n)$ factor and hence does not give a scaling distortion depending solely on ϵ. The loss in the Elkin et al. approach stems from the need to bound the diameter in the recursive construction of the spanning tree. In each level of the construction we may allow only a very small increase as these get multiplied in the bound on the total blowup in the overall diameter. In their original work [EEST05] the increase per level is $\Theta(1/\log n)$, which translates to a multiplicative factor in the distortion. In our case we show that the increase can exponentially decrease along the levels. This indeed guarantees a good blowup in the overall diameter but is awful in terms of the distortion. We apply a new technique for bounding the diameter which allows us to limit the number of levels involved. On the other hand, it is clear that for every value of ϵ there is a limited number of levels for which the distortion requirement imposes new constraints. The proof then proceeds to carefully balance these different arguments. Theorem 1.6 uses essentially the same ideas but in a probabilistic embedding setting.

2. Preliminaries. Consider a finite metric space (X, d) and let $n = |X|$. For any point $x \in X$ and a subset $S \subseteq X$ let $d(x, S) = \min_{s\in S} d(x, s)$. If $P = (S; \bar{S})$ is a partition of X, then $d(x, P) = \max\{d(x, S), d(x, \bar{S})\}$. The *diameter* of X is denoted $\mathrm{diam}(X) = \max_{x,y\in X} d(x, y)$. For a point $x \in X$ and $r \ge 0$, the ball at radius r around x is defined as $B_X(x, r) = \{z \in X | d(x, z) \le r\}$, and the open ball is $B^\circ_X(x, r) = \{z \in X | d(x, z) < r\}$. Given $x \in X$ let $\mathrm{rad}_x(X) = \max_{y\in X} d(x, y)$. Given an edge-weighted graph $G = (X, E, w)$ with $w : E \to \mathbb{R}^+$, let (X, d_X) be the metric space induced from the graph in the usual manner—vertices are associated with points, and distances between points correspond to shortest-path distances in G. If X is clear from the context we may omit the subscript.

An ultrametric (U, d) is a metric space satisfying a strong form of the triangle inequality, that is, for all $x, y, z \in U$, $d(x, z) \le \max\{d(x, y), d(y, z)\}$. The following definition is known to be equivalent to the above definition (see [BLMN05]).

Definition 2.1. *An ultrametric U is a metric space (U, d) whose elements are the leaves of a rooted labeled tree T. Each $v \in T$ is associated a label $\Phi(v) \ge 0$ such that if $u \in T$ is a descendant of v, then $\Phi(u) \le \Phi(v)$ and $\Phi(u) = 0$ iff $u \in U$ is a* *leaf. The distance between leaves $x, y \in U$ is defined as $d(x, y) = \Phi(\mathrm{lca}(x, y))$, where $\mathrm{lca}(x, y)$ is the least common ancestor of x and y in T.*

2.1. Scaling distortion and average distortion. We now prove that a bound of $O(\sqrt{1/\epsilon})$ on the scaling distortion will imply the ℓ_q-distortion bounds as stated in the theorems.

LEMMA 2.2. *If an embedding of an n point metric space has scaling distortion $O(\sqrt{1/\epsilon})$, then it has ℓ_q-distortion:*

- *$O(1)$ for any fixed $q < 2$.*
- *$O(\sqrt{\log n})$ for $q = 2$.*
- *$O(n^{1-2/q})$ for any fixed $q > 2$.*

Proof. Note that for $q = \infty$, taking $\epsilon = 1/n^2$ suggests that all pairs have distortion at most $O(n)$.

By Lemma 1.3 with $\alpha(x) = x^{-1/2}$ it is enough to bound the integral $(\int_{1/n^2}^1 x^{-q/2} dx)^{1/q}$. If $q \neq 2$, then

$$\left(\int_{1/n^2}^1 x^{-q/2} dx\right)^{1/q} = \left(\left[\frac{x^{1-q/2}}{1-q/2}\right]_{1/n^2}^1\right)^{1/q} = \left(\frac{1-n^{q-2}}{1-q/2}\right)^{1/q}.$$

Now for $q < 2$ this is bounded by $(1 - q/2)^{-1/q}$ which is $O(1)$ for any fixed value of q in this range.

For $q > 2$ the integral is $(q/2 - 1)^{-1/q}(n^{q-2} - 1)^{1/q}$. As the term $(q/2 - 1)^{-1/q}$ is $O(1)$ for any fixed value of q in this range the integral is bounded by $O(n^{1-2/q})$.

Finally for $q = 2$ we have

$$\left(\int_{1/n^2}^1 x^{-1} dx\right)^{1/2} = \left([\ln x]_{1/n^2}^1\right)^{1/2} = (2 \ln n)^{1/2}. \qquad \square$$

3. Scaling embedding into an ultrametric. In this section we prove Theorem 1.4. Let (X, d) be a metric space with $n = |X|$ and $\Delta = \mathrm{diam}(X)$. In what follows we will always assume that $\epsilon \geq 1/n^2$, because the distortion bound for $\epsilon = 1/n^2$ holds for all pairs, and the scaling distortion function is monotone. The ultrametric will be represented by a binary tree which is induced by a laminar hierarchical partition of X; each node u corresponds to a subset $X_u \subseteq X$ such that if v, w are the children of u in the ultrametric, then $X_v \cap X_w = \emptyset$, $X_v \cup X_w = X_u$. Furthermore the root r has $X_r = X$ and each leaf corresponds to a singleton.

The high level construction of T is as follows: find a partition P of X into X_1 and $X_2 = X \setminus X_1$; the root of T will be labeled Δ, and its children will be the trees T_1, T_2 formed recursively from the ultrametric trees of X_1 and X_2, respectively. For any $0 < \epsilon < 1$ denote by $B_\epsilon(X)$ the total number of pairs $x, y \in X$ such that $d_T(x, y) > (150/\sqrt{\epsilon})d_X(x, y)$. Note that since the root is labeled by Δ it always holds that $d_T(x, y) \geq d_X(x, y)$, so it remains to bound $B_\epsilon(X)$ by $\epsilon\binom{|X|}{2}$. For a partition $P = (X_1; X_2)$ let $\hat{B}_\epsilon(P) = |\{\{x, y\} \mid x \in X_1 \wedge y \in X_2 \wedge d_X(x, y) < (\sqrt{\epsilon}/150) \cdot \Delta\}|$.

LEMMA 3.1. *For any metric space (X, d) there exists a nontrivial partition*[4] *$P = (X_1; X_2)$ of X such that for any $\epsilon \in (0, 1)$, $\hat{B}_\epsilon(P) \leq \epsilon|X_1||X_2|$.*

[4] A partition $(X_1; X_2)$ is nontrivial if both $X_1, X_2 \neq \emptyset$.

Using this lemma, the proof of the main theorem quickly follows.

Proof of Theorem 1.4. The proof is by induction on the size of X. The base case, where $|X| = 2$, holds because the unique pair realizes the diameter, thus $B_\epsilon(X) = 0$. Assume that for any metric space with $m < n$ points, we can find an ultrametric such that the number of pairs distorted by more than $150/\sqrt{\epsilon}$ is bounded by $\epsilon\binom{m}{2}$. Now consider the metric space (X, d) with $|X| = n$. Let P be the partition $(X_1; X_2)$ guaranteed to exist by Lemma 3.1. By induction,

$$\begin{aligned} B_\epsilon(X) &= \hat{B}_\epsilon(P) + B_\epsilon(X_1) + B_\epsilon(X_2) \\ &\le \epsilon\left(\binom{|X_1|}{2} + \binom{|X_2|}{2} + |X_1| \cdot |X_2|\right) \\ &= \epsilon\binom{|X|}{2}. \end{aligned}$$

This bounds the scaling distortion by $O(1/\sqrt{\epsilon})$ as required, and the consequences of this bound on the ℓ_q-distortion are given in Lemma 2.2. ∎

3.1. Proof of Lemma 3.1. First the partition algorithm is described, then the proof of correctness is done separately for "small" and "large" values of ϵ in Claims 3.3 and 3.4, respectively.

Partition algorithm. Let $u \in X$ be such that $|B^\circ(u, \Delta/2)| \le n/2$; one can always find such a point by considering a pair $u, v \in X$ that realizes the diameter. Let $\hat{\epsilon} = \max\{\epsilon \in (0, 1) \ : \ |B(u, \sqrt{\epsilon}\Delta/4)| \ge \epsilon n\}$ and $\bar{\epsilon} = 32\hat{\epsilon}$; the maximum is indeed obtained because the metric space is finite. Since a ball always contains at least one point we have that $\hat{\epsilon} \ge 1/n$. For all $\epsilon \in (0, 1)$, we have that $B(u, \sqrt{\epsilon}\Delta/4) \subseteq B^\circ(u, \Delta/2)$; by the choice of u this ball contains at most $n/2$ points, thus $\hat{\epsilon} \le 1/2$. Define the intervals $\hat{S} = [\sqrt{\hat{\epsilon}}\Delta/4, \sqrt{\hat{\epsilon}}\Delta/2]$, $S = [(\frac{1}{4} + \frac{1}{25})\sqrt{\hat{\epsilon}}\Delta, (\frac{1}{2} - \frac{1}{25})\sqrt{\hat{\epsilon}}\Delta]$, the length of S, $s = \frac{17}{100}\sqrt{\hat{\epsilon}}\Delta$, and the shell $Q = \{w \ : \ d(u, w) \in \hat{S}\}$. The partition P is defined by carefully choosing a certain $r \in S$ and letting $X_1 = B(u, r)$ and $X_2 = X \setminus X_1$. The following property will be useful,

PROPOSITION 3.2. $|B(u, \sqrt{\hat{\epsilon}}\Delta/2)| \le 4\hat{\epsilon}n$.

Proof. There are two cases. If $\hat{\epsilon} \le 1/4$, then $|B(u, \sqrt{\hat{\epsilon}}\Delta/2)| = |B(u, \sqrt{4\hat{\epsilon}}\Delta/4)| \le 4\hat{\epsilon}n$ (otherwise it is a contradiction to the maximality of $\hat{\epsilon}$). In the other case $\hat{\epsilon} \in (1/4, 1)$, but now $|B(u, \sqrt{\hat{\epsilon}}\Delta/2)| \le |B^\circ(u, \Delta/2)| \le n/2 \le 2\hat{\epsilon}n$. ∎

We will now show that a certain choice of $r \in S$ will produce a partition that satisfies the condition of Lemma 3.1 for all $\epsilon \in (0, \bar{\epsilon}]$. For any $r \in S$ and $\epsilon \le \bar{\epsilon}$ let $S_r(\epsilon) = (r - \sqrt{\epsilon}\Delta/150, r + \sqrt{\epsilon}\Delta/150)$, $s(\epsilon) = \sqrt{\epsilon}\Delta/75$, and let $Q_r(\epsilon) = \{w \ : \ d(u, w) \in S_r(\epsilon)\}$. Notice that for any $r \in S$ and any $\epsilon \le \bar{\epsilon}$, $S_r(\epsilon) \subseteq \hat{S}$. Define that property $A_r(\epsilon)$ holds if the shell $Q_r(\epsilon)$ has sufficiently small cardinality, which will imply that cutting at radius r is "good" for ϵ. Formally,

$$A_r(\epsilon) \text{ holds iff } |Q_r(\epsilon)| \le \sqrt{\epsilon \cdot \hat{\epsilon}/2} \cdot n. \tag{3.1}$$

Note that the triangle inequality suggests that only pairs $\{x, y\}$ such that $x, y \in Q_r(\epsilon)$ may contribute to $\hat{B}_\epsilon(P)$. Indeed, assume that $x \in X_1$, $y \in X_2$ and withost loss of generality (w.l.o.g.) $y \notin Q_r(\epsilon)$; then $d(x, y) \ge d(u, y) - d(x, u) \ge r + \sqrt{\epsilon}\Delta/150 - r = \sqrt{\epsilon}\Delta/150$, so by definition $\{x, y\} \notin \hat{B}_\epsilon(P)$. The other case when $x \notin Q_r(\epsilon)$ is symmetric. Observe that $r \ge \sqrt{\hat{\epsilon}}\Delta/4$, and thus by the definition of $\hat{\epsilon}$

$$|X_1| \ge |B(u, \sqrt{\hat{\epsilon}}\Delta/4)| \ge \hat{\epsilon}n. \tag{3.2}$$

We also have that $r < \sqrt{\hat\epsilon}\Delta/2$, so that $|X_1| \le |B(u, \sqrt{\hat\epsilon}\Delta/2)| \le n/2$ (by the choice of u), so that

$$|X_2| \ge n/2. \tag{3.3}$$

We conclude that if $A_r(\epsilon)$ holds, then by (3.1), (3.2), and (3.3),

$$\hat{B}_\epsilon(P) \le |Q_r(\epsilon)|^2 \le \epsilon \cdot \hat\epsilon n^2/2 \le \epsilon |X_1| n/2 \le \epsilon |X_1||X_2|,$$

that is, the condition of Lemma 3.1 is satisfied for ϵ. Hence for $\epsilon \in (0, \bar\epsilon]$ the following is sufficient.

Claim 3.3. *There exists some $r \in S$ such that property $A_r(\epsilon)$ holds for all $\epsilon \in (0, \bar\epsilon]$ simultaneously.*

Proof. The proof is based on the following iterative process that greedily deletes the "worst" interval in S. Initially, let $I_0 = S$, and $j = 1$:

1. If for all $r \in I_{j-1}$ and for all $\epsilon \le \bar\epsilon$ property $A_r(\epsilon)$ holds, then set $t = j - 1$, stop the iterative process, and output I_t.
2. Let $\mathcal{S}_j = \{S_r(\epsilon) \;:\; r \in I_{j-1}, \epsilon \le \bar\epsilon, \neg A_r(\epsilon)\}$. We greedily remove the interval that has maximal ϵ. Formally, let r_j, ϵ_j be parameters such that $S_{r_j}(\epsilon_j) \in \mathcal{S}_j$ and $\epsilon_j = \max\{\epsilon \;:\; \exists S_r(\epsilon) \in \mathcal{S}_j\}$.
3. Set $I_j = I_{j-1} \setminus S_{r_j}(\epsilon_j)$, set $j = j + 1$, and goto 1.

Note that this process can be computed in polynomial time, using a simple discretization of ϵ and r. For instance, as we would like to determine for a fixed ϵ if there is an r such that $|Q_r(\epsilon)|$ is too large, we need to scan all the intervals of length $s_r(\epsilon)$. But we claim that it suffices to scan those intervals that start or end at a point, that is, at $d(u, w)$ for some $w \in Q$ (all the other intervals will not contain more points). One can then do a simple binary search on the value of ϵ to find the largest (we only need to consider polynomially many values, those values for which $\epsilon \cdot \hat\epsilon n^2/2$ is an integer).

We now argue that $I_t \ne \emptyset$ and hence an appropriate value $r \in S$ can be found. First we will show that any point $a \in \hat{S}$ can be covered by at most two intervals $S_{r_j}(\epsilon_j)$, $S_{r_i}(\epsilon_i)$ for some $1 \le j < i \le t$. This holds because once a is covered from the left by $S_{r_j}(\epsilon_j)$ (that is, $r_j \le a$), then this interval is removed in step 3. from I_j. By maximality of ϵ_j, for any $j < i \le t$ we have that $s(\epsilon_i) \le s(\epsilon_j)$, so as $r_i \in I_j$ it must be that the interval $S_{r_i}(\epsilon_i)$ covers a from the right (that is, $r_i \ge a$), and no other interval will cover a in the remainder of the process.

Observe that $S_{r_j}(\epsilon_j) \subseteq \hat{S}$ for any $0 \le j \le t$. This suggests that any $x \in Q$ appears in at most two sets $Q_{r_j}(\epsilon_j)$, $Q_{r_i}(\epsilon_i)$. From this and Proposition 3.2,

$$\sum_{j=1}^{t} |Q_{r_j}(\epsilon_j)| \le 2|Q| \le 8\hat\epsilon n. \tag{3.4}$$

Recall that since $A_{r_j}(\epsilon_j)$ does not hold for any $1 \le j \le t$, then

$$\sum_{j=1}^{t} |Q_{r_j}(\epsilon_j)| \ge \sqrt{\hat\epsilon/2} \cdot n \sum_{j=1}^{t} \sqrt{\epsilon_j}, \tag{3.5}$$

and combining (3.4) and (3.5) yields

$$\sum_{j=1}^{t} \sqrt{\epsilon_j} \le 12\sqrt{\hat\epsilon}.$$

Finally we can bound the total length of the "bad" intervals chosen by the process by the definition of $s(\epsilon)$,

$$\sum_{j=1}^{t} |S_{r_j}(\epsilon_j)| = \sum_{j=1}^{t} s(\epsilon_j) \le \sum_{j=1}^{t} \sqrt{\epsilon_j}\Delta/75 \le 12/75 \cdot \sqrt{\hat{\epsilon}}\Delta = 16/100 \cdot \sqrt{\hat{\epsilon}}\Delta.$$

Since $|I_0| = s = 17/100 \cdot \sqrt{\hat{\epsilon}}\Delta$ it is impossible that the entire interval I_0 was removed; therefore $I_t \neq \emptyset$, and actually any $r \in I_t$ satisfies the condition of Claim 3.3. $\square$

Next we show that in fact any choice of $r \in S$ will produce a partition that satisfies Lemma 3.1 for all $\epsilon \in (\bar{\epsilon}, 1)$.

CLAIM 3.4. *If $\epsilon \in (\bar{\epsilon}, 1)$, $r \in S$, and $P = (B(u,r); X \setminus B(u,r))$, then $\hat{B}_\epsilon(P) < \epsilon|X_1||X_2|$.*

Proof. Note that only pairs $\{x,y\}$ such that $x \in X_1$ and $y \in B(u, r+\sqrt{\epsilon}\Delta/16) \cap X_2$ can be distorted by more than $16\sqrt{1/\epsilon}$ and hence may be counted in $\hat{B}_\epsilon(P)$, so

$$|\hat{B}_\epsilon(P)| < |X_1| \cdot |B(u, r + \sqrt{\epsilon}\Delta/16)|. \tag{3.6}$$

Since $\sqrt{\hat{\epsilon}} \le \sqrt{\epsilon/2}/4$ and $r < \sqrt{\hat{\epsilon}}\Delta/2 \le \sqrt{\epsilon/2}\Delta/8$, then

$$|B(u, r + \sqrt{\epsilon}\Delta/16)| \le \left| B\left(u, \sqrt{\epsilon/2}\left(\frac{1}{8} + \frac{1}{8}\right)\Delta\right)\right| = |B(u, \sqrt{\epsilon/2}\Delta/4)| < \epsilon n/2, \tag{3.7}$$

where the last inequality used that $\epsilon/2 > \hat{\epsilon}$ and the maximality of $\hat{\epsilon}$. Plugging (3.7) into (3.6) and using (3.3) it follows that $\hat{B}_\epsilon(P) < \epsilon|X_1| \cdot |X_2|$, as required. $\square$

4. Scaling embedding into a spanning tree. Here we extended the techniques of the previous section, in conjunction with the constructions of [EEST05], to achieve the following.

THEOREM 1.5 (restated). *Any weighted graph of size n contains a spanning tree with scaling distortion $O(\sqrt{1/\epsilon})$. In particular, its ℓ_q-distortion is $O(1)$ for any fixed $1 \le q < 2$, $O(\sqrt{\log n})$ for $q = 2$, and $O(n^{1-2/q})$ for any fixed $2 < q \le \infty$.*

Given a graph, the spanning tree is created by recursively partitioning the graph using a *hierarchical star-partition*. The algorithm has three components, with the following high level description:

1. A decomposition algorithm that creates a single cluster. The decomposition algorithm is similar in spirit to the decomposition algorithm used in the previous section for metric spaces. We will later explain the main differences.

2. A star-partition algorithm. This algorithm partitions a graph X into a central ball X_0 with center x_0 and a set of cones $X_1, \ldots, X_m$ and also outputs a set of edges of the graph $\{y_1, x_1\}, \ldots, \{y_m, x_m\}$ that connect each cone-set X_i to the central ball X_0 by the edge $\{x_i, y_i\}$, where $x_i \in X_i$ and $y_i \in X_0$. The central ball is created by invoking the decomposition algorithm with a center x_0 to obtain a cluster whose radius is in the range $[(1/2)\mathrm{rad}_{x_0}(X), (5/8)\mathrm{rad}_{x_0}(X)]$. Each cone-set X_i is created by invoking the decomposition algorithm on a certain "cone-metric" to be defined in what follows. Informally, a ball in the cone-metric around x_i with radius r is the set of all points x such that $d(x_0, x_i) + d(x_i, x) - d(x_0, x) \le r$. Hence each cone X_i is a ball whose center is x_i in some appropriately defined cone-metric. The radius of each ball in the cone-metric is chosen to be $\approx \tau^k \mathrm{rad}_{x_0}(X)$, where $\tau < 1$ is some fixed constant and k is the depth of the recursion since the *last reset cluster*. Unfortunately, at some stage the radius may be too small for the decompose algorithm to preform well enough. In such cases we must reset the parameters that govern the radius of

the cones. (In the next item, we will define more accurately how the recursion is performed and when this parameter of a cluster may be reset.) The main property of this star decomposition is that for any point $x \in X_i$, the distance to the center x_0 does not increase by more than r. In the paper [EEST05] the radius r was chosen to be $\approx \mathrm{rad}_{x_0}(X)/\log n$, thus the total radius increase over the $O(\log n)$ levels of recursion was $O(1)$. We cannot allow the cone radius to depend on n, because this translates to a loss in the distortion, so we use a different method to guarantee $O(1)$ radius increase.

3. Recursive application of the star-partition. As mentioned in the previous item, the radius of the balls in the cone-metric are exponentially decreasing. However, at certain stages in the recursion, the cone radius becomes too small and the parameters governing the cone radius must be reset. Clusters in which the parameters need to be restarted are called *reset clusters*. The two parameters that are associated with a reset cluster X are $n = |X|$ and $\Lambda = \mathrm{rad}(X)$. Specifically, a cluster is called a reset cluster if its size relative to the size of the last reset cluster is larger than some constant times its radius relative to radius of the last reset cluster. In that case n and Λ are updated to the values of the current cluster. This implies that reset clusters have small diameter, hence their total contribution to the increase of radius is small. Moreover, resetting the parameters allows the decompose algorithm to continue to produce the clusters with the necessary properties to obtain the desired scaling distortion. Using resets, the algorithm can continue recursively in this fashion until the spanning tree is formed.

Decompose algorithm. The decompose algorithm receives as input several parameters. First it obtains a pseudometric space (W, ρ) and point u (for the central ball this is just the shortest-paths metric, while for cones, this pseudometric is the so-called cone-metric which will be formally defined in what follows). The goal of the decompose algorithm is to partition W into a cluster which is a ball $Z = B_{(W,\rho)}(u, r)$ and $\bar{Z} = W \setminus Z$.

Informally, this partition $P = (Z; \bar{Z})$ is carefully chosen to maintain the scaling property: for every ϵ, the number of pairs whose distortion is too large is "small enough" (an ϵ-fraction of the separated pairs). Let $\hat{\Lambda}$ be a parameter corresponding to the radius of the cluster over which the star-partition is performed. Pairs that are separated "close" to the partition may risk the possibility of being at distance $\Theta(\hat{\Lambda})$ in the constructed spanning tree. One of the technical difficulties in the graph setting is that unlike the metric case, a pair $\{u, v\}$ can suffer distortion by a partition that *does not separate* u from v: it suffices that the partition cuts the shortest path from u to v. For certain values of ϵ, we denote by $\hat{B}_\epsilon(P)$ the number of pairs that may be distorted by at least $\Omega(\sqrt{1/\epsilon})$ if the distance between them will grow to $\hat{\Lambda}$. Using a decomposition which iteratively deletes bad intervals, similar in spirit to the one used in Lemma 3.1, we expect the number of "bad" pairs for a specific value of ϵ to be at most ϵ-fraction of the total possible number of separated pairs. However, if we insist that this property holds true for all ϵ we cannot maintain a small enough bound on the maximum value for the radius r. Roughly speaking, r must have a possible range of size $\approx \sqrt{\epsilon}\hat{\Lambda}$ in order to succeed for ϵ. Since the size of this range determines the amount of increase in the radius of the cluster, we would like to be able to bound it. Therefore, we keep another parameter, denoted $\epsilon_{\lim} = \epsilon_{\lim}(W)$ (we will often omit W when it is clear from context; also note that this parameter may be larger than 1). That is, the partition P will be good only for those values of ϵ satisfying $\epsilon \le \epsilon_{\lim}$. This bound on the range of ϵ values actually allows us to give a stronger bound than in Lemma 3.1 on the number of "bad" pairs, which improves by a factor of the additional new parameter β.

The radius r of the ball is controlled by the radius of the cluster $\hat{\Lambda}$ and by two new parameters θ and $\alpha \approx \sqrt{\epsilon_{\text{lim}}}$. The guarantee is that $r \in [\theta\hat{\Lambda}, (\theta + \alpha)\hat{\Lambda}]$. For the central ball of the star-partition θ is fixed to $1/2$ and for the star's cones θ is fixed to 0. Indeed, as indicated above, the value of ϵ_{lim} determines the increase in the radius of the cluster by setting the value for α, which gives enough range in the choice of radius to succeed for all $\epsilon \leq \epsilon_{\text{lim}}$.

Note that there are two conflicting constraints here. On the one hand we want ϵ_{lim} to be large so that the partition of the current level will be successful for many values of ϵ. On the other hand we need that the total radius increase over all levels will be bounded, so this level must "pay its toll" and allow only a small increase in the radius, which immediately translate to an upper bound on ϵ_{lim}. As it turns out, setting $\epsilon_{\text{lim}} = |W|/(n \cdot \beta)$ will satisfy both requirements simultaneously. It will decrease in a geometric manner as long as there is no reset, which is very useful for the bound on the total radius increase. On the other hand it is still large enough for controlling the number of distorted pairs, because for $\epsilon > \epsilon_{\text{lim}}$, the total number of pairs in W, which is $\approx |W|^2$, is a small enough fraction of n^2, so we might as well consider them all as distorted.

Let us explain now how the decompose algorithm will be used within our overall scheme. A useful property is that the radius of the clusters in the hierarchical star decomposition decreases geometrically. The parameter β is chosen to be polynomial in the ratio between the current radius to that of the last reset cluster, so it is bounded by μ^k, where $\mu < 1$ is some fixed constant and k is the depth of the recursion from the last reset cluster. There will be three types of ways to count distorted pairs. Our decompose algorithm generates a parameter $\bar{\epsilon}$ for each cluster it cuts, which distinguishes small and large values of ϵ, similarly to the distinction in the proof of Lemma 3.1.

1. For each $\epsilon < \bar{\epsilon}$ the notation $\hat{B}_\epsilon(P)$ for a partition $P = (Z; \hat{Z})$ will stand for the number of pairs that may be distorted by invoking the partition P. Informally it consists of all the pairs $\{u, v\}$ such that both u, v are of distance less than $\approx \sqrt{\epsilon}\hat{\Lambda}$ from the cut (in the metric (W, ρ)). The property obtained by the decompose algorithm is that $\hat{B}_\epsilon(P)$ is at most $O(\epsilon|Z| \cdot (n - |Z|) \cdot \beta)$.
2. For $\bar{\epsilon} \leq \epsilon \leq \epsilon_{\text{lim}}$ and a partition $P = (Z; \hat{Z})$, we use a different counting argument. The proof for the metric case does not suffice for the "large" values of ϵ in the spanning tree case, because in the latter case there are potentially many more pairs that are in danger of being distorted (those whose shortest path is cut by the partition). This is why we require a different argument for this range of the parameter ϵ: If a point u is close enough ($\approx \sqrt{\epsilon}\hat{\Lambda}$) to the cut, we simply throw away all pairs $\{u, v\}$ where v is $\approx \sqrt{\epsilon}\hat{\Lambda}$ close to u (in the *induced* metric on the cluster, not the cone-metric). These are all the pairs that can be distorted by more than $O(1/\sqrt{\epsilon})$. Our decompose scheme will guarantee that there are only $\approx \epsilon n$ such points for any $u \in W$. Furthermore, it will be shown that this throwing is done only once throughout the whole recursion for a point $u \in V(G)$ and a fixed $\epsilon > 0$. Let $\bar{B}_\epsilon(G)$ (defined just below), denote all the pairs counted in this way.
3. For ϵ that is larger than ϵ_{lim}, we show that the number of points in the current cluster is less than an ϵ fraction of the number of points in the last

reset cluster; hence we can discard all the pairs in such clusters and the total number of all such discarded pairs is small.

We now turn to the formal description of the algorithm and its analysis. Assume a cluster X is partitioned to $X_0, X_1, \ldots, X_m$ by invoking the decompose algorithm that generates partitions $P_0, \ldots, P_{m-1}$, where $P_i = (X_i, Y_i)$ and $Y_i = X \setminus \cup_{0\le j\le i} X_j$. Then define recursively

$$B_\epsilon(X) = \sum_{j=0}^{m-1} \hat{B}_\epsilon(P_i) + \sum_{j=0}^{m} B_\epsilon(X_i) \tag{4.1}$$

(where $\hat{B}_\epsilon(P_i)$ is defined in Lemma 4.1). The base case is when $|X| = 1$, or when $\epsilon > \epsilon_{\lim}$ in such a case $B_\epsilon(X) = \binom{|X|}{2}$. Note that the definition of $B_\epsilon(X)$ may ignore the pairs in $\bar{B}_\epsilon(G) = \sum_{x\in V(G)} \bar{B}_\epsilon(x)$ (where $\bar{B}_\epsilon(x)$ is defined in Lemma 4.1). Indeed those pairs will be accounted for separately.

We will make use of the following predefined constants: $c = e+1$, $c' = 2e+1$, $\hat{c} = 44$, and $C = 8\sqrt{c\cdot\hat{c}}$. Finally, the distortion is given by $600C\cdot c'$. The exact properties of the decomposition algorithm are captured by the following lemma.

Lemma 4.1. *Given a (pseudo) metric space (W,ρ), a graph metric d_W on W, a point $u \in W$, and parameters $n \ge |W|$, $\hat{\Lambda} > 0$, $0 < \beta < 1/\hat{c}$, and $\theta \in \{0, 1/2\}$, there exists an algorithm $\mathtt{decompose}((W,\rho), u, \hat{\Lambda}, \theta, n, \epsilon_{\lim}, \beta)$, where $\epsilon_{\lim} \ge \frac{|W|}{\beta\cdot n}$, that computes a partition $P = (Z;\bar{Z})$ of W such that $Z = B_{(W,\rho)}(u,r)$ and $r/\hat{\Lambda} \in [\theta, \theta+\alpha]$, where $\alpha = \sqrt{\epsilon_{\lim}}/C$. It also returns a parameter $\bar{\epsilon} > 0$. Let $S_\epsilon(P) = B_{(W,\rho)}(u, r + \frac{\sqrt{\epsilon}\cdot\hat{\Lambda}}{150C}) \setminus B_{(W,\rho)}(u, r - \frac{\sqrt{\epsilon}\cdot\hat{\Lambda}}{150C})$ and for $\epsilon \le \bar{\epsilon}$ let $\hat{B}_\epsilon(P) = |S_\epsilon(P)|^2$ (for $\epsilon > \bar{\epsilon}$ we set $\hat{B}_\epsilon(P) = 0$). The partition has the property that for any $\epsilon \in (0, \bar{\epsilon}]$,*

$$\hat{B}_\epsilon(P) \le \epsilon |Z|\cdot(n - |Z|)\cdot\beta.$$

For any $\epsilon \in [\bar{\epsilon}, \epsilon_{\lim}]$ and for any $x \in S_\epsilon(P)$ for which $\bar{B}_\epsilon(x)$ is not yet defined,

$$\bar{B}_\epsilon(x) := \left| B_{(W,d_W)}\left(x, \frac{\sqrt{\epsilon}\cdot\hat{\Lambda}}{150C}\right)\right| \le \epsilon n/8. \tag{4.2}$$

We defer the proof of this technical lemma to the end of the section.

Star-partition algorithm. Consider a cluster X with center x_0 and parameters n, Λ. Recall that parameters n, Λ are the number of points and the radius (respectively) of the last reset cluster. A star-partition partitions X into a central ball X_0, and cone-sets $X_1, \ldots, X_m$ and edges $\{y_1, x_1\}, \ldots, \{y_m, x_m\}$, and the value m is determined by the star-partition algorithm when no more cones are required. Each cone-set X_i is connected to X_0 by the edge $\{y_i, x_i\}, y_i \in X_0, x_i \in X_i$. Denote by $P_0 = (X_0; X\setminus X_0)$ the partition creating the central ball X_0 and by $\{P_i\}_{i=1}^m$ the partitions creating the cones, where $P_i = (X_i; X\setminus(\cup_{0\le j\le i} X_j))$. In order to create the cone-set X_i use the decompose algorithm on the cone-metric $\ell_{x_i}^{x_0}$ defined below. We refer the reader to [EEST05, ABN08] for the intuition behind this definition.

Definition 4.2 (cone-metric). *Given a graph $G = (X,E)$ with shortest path metric d, a set $Y \subset X$, $x \in X$, $y \in Y$ define the* cone-metric *$\ell_y^x : Y^2 \to \mathbb{R}^+$ as $\ell_y^x(u,v) = |(d(x,u) - d_Y(y,u)) - (d(x,v) - d_Y(y,v))|$, where d_Y is the metric induced by shortest paths in the subgraph $(Y, E[Y])$.*

$T=$ `hierarchical-star-partition`(X,x,n,Λ):

1. If $|X|=1$ set $T=X$ and stop.
2. $(X_0,\ldots,X_m,\{y_1,x_1\},\ldots,\{y_m,x_m\})=$ `star-partition`(X,x,n,Λ);
3. For each $i\in[0,1,\ldots,m]$:
   (a) If $\frac{|X_i|}{n}<c^{\frac{\mathrm{rad}_{x_i}(X_i)}{\Lambda}}$ then $T_i=$ `hierarchical-star-partition` (X_i,x_i,n,Λ);
   (b) Otherwise, set X_i to be a **reset cluster**, and $T_i=$ `hierarchical-star-partition`$(X_i,x_i,|X_i|,\mathrm{rad}_{x_i}(X_i))$;
4. Let T be the tree formed by connecting T_0 with T_i using edge $\{y_i,x_i\}$ for each $i\in[1,\ldots,m]$;

FIG. 1. *Hierarchical star-partition algorithm.*

$(X_0,\ldots,X_m,\{y_1,x_1\},\ldots,\{y_m,x_m\})=$ `star-partition`(X,x_0,n,Λ):

1. Set $i=0$; $\beta=\frac{1}{\hat{c}}(\frac{\mathrm{rad}_{x_0}(X)}{\Lambda})^{1/4}$; $\epsilon_{\lim}=|X|/(\beta n)$; $\hat{\Lambda}=\mathrm{rad}_{x_0}(X)$;
2. $(X_0,Y_0)=$ `decompose`$((X,d),x_0,\hat{\Lambda},1/2,n,\epsilon_{\lim},\beta)$; (`decompose` is given by Lemma 4.1);
3. If $Y_i=\emptyset$ set $m=i$ and stop; Otherwise, set $i=i+1$;
4. Let $\{x_i,y_i\}$ be an edge in E such that $y_i\in X_0,x_i\in Y_{i-1}$, with y_i on the shortest path from x_i to x_0;
5. Let $\ell=\ell_{x_i}^{x_0}$ be cone-metric on the graph induced by $X_0\cup Y_{i-1}$, the set Y_{i-1} and the points x_0,x_i;
6. $(X_i,Y_i)=$ `decompose`$((Y_{i-1},\ell),x_i,\hat{\Lambda},0,n,\epsilon_{\lim},\beta)$;
7. goto 3;

FIG. 2. *Star-partition algorithm.*

Note that the cone-metric ℓ is in fact a pseudometric (it could be that $\ell_y^x(u,v)=0$ for $u\neq v$). Also note that

$$B_{(Y,\ell_y^x)}(y,r)=\{v\in Y|d(x,y)+d_Y(y,v)-d(x,v)\le r\}. \tag{4.3}$$

The following fact will be useful. For all $u,v\in Y$, since $d(u,v)\le d_Y(u,v)$ and by the triangle inequality,

$$\ell_y^x(u,v)\le|d(x,u)-d(x,v)|+|d_Y(y,u)-d_Y(y,v)|\le 2d_Y(u,v). \tag{4.4}$$

Hierarchical star-partition algorithm. Given a graph $G=(V,E,w)$, the metric d induced by the graph is the shortest path metric. The construction of the spanning tree for G is done by choosing some $x\in V$, setting V as a reset cluster and calling `hierarchical-star-partition`$(V,x,|V|,\mathrm{rad}_x(V))$; see Figures 1 and 2.

4.1. Algorithm analysis. The hierarchical star-partition of $G=(X,E,w)$ naturally induces a laminar family $\mathcal{F}\subseteq 2^X$. Let $\mathcal{G}$ be the rooted *construction tree* whose nodes are sets in $\mathcal{F}$, and $F\in\mathcal{F}$ is a parent of $F'\in\mathcal{F}$ if F' is a cluster formed by the partition of F. The root of $\mathcal{G}$ corresponds to X and the leaves to the singletons. Note that every leaf is a reset cluster (because its radius is 0). Observe that the spanning tree T obtained by our hierarchical star decomposition has the property that every $F\in\mathcal{F}$ corresponds to a connected subtree $T[F]$ of T. Let $\mathcal{R}\subseteq\mathcal{F}$ be the set of all reset clusters. For each $F\in\mathcal{F}$, let $\mathcal{G}_F$ be the subtree of the construction tree $\mathcal{G}$ rooted at F that contains all the nodes X whose path to F (excluding F and X) contains no node in $\mathcal{R}$. In other words, $\mathcal{G}_F$ is the tree rooted at F whose leaves are reset clusters, such that all inner nodes (except maybe for F itself) are not reset clusters. For $F \in \mathcal{F}$ let $\mathcal{R}(F) \subseteq \mathcal{R}$ be the set of leaves of $\mathcal{G}_F$. For $Y \in \mathcal{R}$ and $F \in \mathcal{G}_Y$ denote by $L_Y(F) = L(F) = d_{\mathcal{G}}(F, Y)$ the distance in the construction tree from F to Y; this is the number of recursion levels since the last reset occurred, and naturally $L(Y) = 0$ for all reset clusters (we shall omit the subscript when Y is clear from context).

In what follows the parameter α for each cluster is defined as in Lemma 4.1, in particular, for a cluster X, $\alpha(X) = \frac{1}{C} \cdot \sqrt{\frac{\hat{c} \cdot |X|}{n}} \cdot (\frac{\Lambda}{\mathrm{rad}(X)})^{1/8}$, where n and Λ are the parameters of the last reset cluster. Also we use the following convention in our notation: whenever X is a cluster in $\mathcal{G}$ with center point x_0 with respect to which the star-partition of X has been constructed, we define $\mathrm{rad}(X) = \mathrm{rad}_{x_0}(X)$.

Steiner points. For the sake of analysis, we will imagine that whenever the star-partition algorithm on (X, x_0) adds an edge $\{y_i, x_i\}$ between X_0 and X_i, we also add an imaginary Steiner point y_i' on the edge $\{y_i, x_i\}$ such that $d_X(x_0, y_i') = \mathrm{rad}(X_0)$. The reason is that the analysis of the total radius increase is simplified if we can provide a bound on the radius increase of a single iteration of the star-partition algorithm of the form

$$\mathrm{rad}(X_0) + d_X(y_i', x_i) + \mathrm{rad}(X_i) \le \mathrm{rad}(X) \cdot (1 + \alpha), \tag{4.5}$$

rather than the bound of $d_X(x_0, x_i) + \mathrm{rad}(X_i) \le \mathrm{rad}(X) \cdot (1 + \alpha)$ that was shown in [EEST05]. To this end, adding this y_i' at a distance of $\mathrm{rad}(X_0)$ from x_0 will enable us to claim that $d_X(x_0, x_i) = d_X(x_0, y_i') + d_X(y_i', x_i) = \mathrm{rad}(X_0) + d_X(y_i', x_i)$, which yields (4.5). Note that these Steiner points will in fact be leaves in the graph induced on X_0, and since the hierarchical star-partition algorithm maintains connectivity, the edge $\{y_i, y_i'\}$ will be in the final tree T. Thus we may remove from T all the Steiner points that we added, without changing any distance. We stress that the Steiner points will play no role whatsoever in the decomposition algorithm nor in the decision for setting a cluster as a reset cluster. That is, whenever computing the cardinality of a set of vertices, we do not count the Steiner points.

We start our analysis by showing the following claim (recall that for $Y \subseteq X$, d_Y is the shortest path metric induced on the subgraph Y).

CLAIM 4.3. *For any cluster X, $x_0 \in X$, $j > 0$ let $Y_{j-1} \subseteq X$ be the unassigned points of X after creating j clusters $X_0, \ldots, X_{j-1}$ using the star-partition algorithm; then for any $z \in Y_{j-1}$ all the shortest paths from z to x_0 are fully contained in $Y_{j-1} \cup X_0$, in particular*

$$d_{Y_{j-1} \cup X_0}(x_0, z) = d_X(x_0, z).$$

Proof. Let $\hat{\Lambda} = \mathrm{rad}_{x_0}(X)$. Let $P_{x_0,z}$ be a shortest path from x_0 to z in X, and seeking contradiction assume that $P_{x_0,z} \not\subseteq Y_{j-1} \cup X_0$. Let $1 \le i \le j-1$ be the minimal such that there exists $u \in P_{x_0,z}$ and $u \in X_i$. Recall that x_i is the center of the cone X_i, and let $r_i \in [0, \alpha\hat{\Lambda}]$ be the radius chosen in Lemma 4.1 when creating X_i. Since $u \in X_i = B_{(X_0 \cup Y_{i-1}, \ell_{x_i}^{x_0})}(x_i, r_i)$, by (4.3) it must be that in the metric $d' = d_{X_0 \cup Y_{i-1}}$

$$d'(x_0, x_i) + d_{Y_{i-1}}(x_i, u) \le d'(x_0, u) + r_i. \tag{4.6}$$

Since u lies on a shortest path from z to x_0, the minimality of i suggests that this shortest path is fully contained in $Y_{i-1} \cup X_0$, thus

$$d'(x_0, z) = d'(x_0, u) + d'(u, z). \tag{4.7}$$

Also note that the shortest path from u to z cannot intersect X_0, because if $a \in P_{u,z} \cap X_0$, then by (4.7), $d'(x_0,z) \le d'(x_0,a) + d'(a,z) < d'(x_0,u) + d'(u,z) = d'(x_0,z)$, which is a contradiction. Thus we obtain that

$$d_{Y_{i-1}}(u,z) = d'(u,z). \tag{4.8}$$

Finally, combining (4.7), (4.6), and (4.8) we conclude that

$$\begin{aligned} d'(x_0,z) + r_i = d'(u,z) + d'(x_0,u) + r_i &\ge d_{Y_{i-1}}(u,z) + d_{Y_{i-1}}(u,x_i) + d'(x_i,x_0) \\ &\ge d_{Y_{i-1}}(z,x_i) + d'(x_i,x_0); \end{aligned}$$

hence z should in fact be in X_i, contradiction. ∎

Next we bound the radius increase due to a single iteration of the star-partition algorithm.

Claim 4.4. *For any cluster X that is partitioned by the star-partition algorithm to $X_0, X_1, \ldots, X_m$ and any $1 \le i \le m$,*

$$\mathrm{rad}(X_0) + d_X(y_i', x_i) + \mathrm{rad}(X_i) \le (1+\alpha)\mathrm{rad}(X).$$

Proof. Fix some integer $i \ge 1$ and consider $z \in X_i$ such that $\mathrm{rad}(X_i) = d_{X_i}(x_i,z) = d_{Y_{i-1}}(x_i,z)$, where the last equality holds since $z \in X_i$. It must be that all the points in Y_{i-1} that are closer than z to x_i (in the cone-metric), in particular those on the shortest paths (in Y_{i-1}) from x_i to z, are also contained in X_i. By Claim 4.3 we know that $d_X(x_0,z) = d_{X_0 \cup Y_{i-1}}(x_0,z)$. By the algorithm in Lemma 4.1, the radius of the cone X_i is selected from the interval $[0, \alpha \cdot \mathrm{rad}(X)]$, so that $z \in B_{(X_0 \cup Y_{i-1}, \ell_{x_i}^{x_0})}(x_i, \alpha \cdot \mathrm{rad}(X))$. By (4.3) this means that in the metric $d' = d_{X_0 \cup Y_{i-1}}$

$$d'(x_0,x_i) + d_{Y_{i-1}}(x_i,z) \le d'(x_0,z) + \alpha \cdot \mathrm{rad}(X). \tag{4.9}$$

Since y_i' is on the shortest path from x_0 to x_i, and at distance $\mathrm{rad}(X_0)$ from x_0, we get by Claim 4.3 that

$$\mathrm{rad}(X_0) + d_X(y_i', x_i) = d_X(x_0, y_i') + d_X(y_i', x_i) = d_X(x_0, x_i) = d'(x_0, x_i). \tag{4.10}$$

Finally, combining (4.9) and (4.10) it follows that

$$\begin{aligned} \mathrm{rad}(X_0) + d_X(y_i', x_i) + \mathrm{rad}(X_i) &= d'(x_0,x_i) + d_{Y_{i-1}}(x_i,z) \\ &\le d'(x_0,z) + \alpha \cdot \mathrm{rad}(X) \\ &= d_X(x_0,z) + \alpha \cdot \mathrm{rad}(X) \\ &\le (1+\alpha) \cdot \mathrm{rad}(X). \qquad ∎ \end{aligned}$$

Corollary 4.5. *For any cluster X that is partitioned by the algorithm to $X_0, X_1, \ldots, X_m$ and any $0 \le i \le m$,*

$$\mathrm{rad}(X_i) \le (1/2+\alpha)\mathrm{rad}(X).$$

Proof. This is immediate for $i = 0$ since $\theta = 1/2$ and the radius of the cluster X_0 is in the interval $[\theta \cdot \mathrm{rad}(X), (\theta+\alpha) \cdot \mathrm{rad}(X)]$ by the definition in Lemma 4.1. For all $1 \le i \le m$, since $\mathrm{rad}(X_0) \ge \mathrm{rad}(X)/2$ we get using Claim 4.4 that

$$\mathrm{rad}(X_i) \le (1+\alpha)\mathrm{rad}(X) - \mathrm{rad}(X_0) \le (1/2+\alpha)\mathrm{rad}(X). \qquad ∎$$

The parameter α is the bound on the *radius increase* created by the star partition. We would like to show that as long as there is no reset, this parameter decreases exponentially fast.

CLAIM 4.6. *Fix some $Y \in \mathcal{R}$. Let $X \in \mathcal{G}_Y \setminus \mathcal{R}(Y)$ with $L_Y(X) = t$. Then the following hold:*

- $\mathrm{rad}(X) \le (\frac{5}{8})^t \mathrm{rad}(Y)$.
- $\alpha = \alpha(X) \le \frac{1}{8}(\frac{7}{8})^t$.

Proof. The first statement is proved by induction on $t = L_Y(X)$. The base case for $t = 0$ implies $X = Y$, so it is trivial. Assume it holds for $t-1$, and to prove the inductive step, it is sufficient to show that when X is partitioned to $X_0, X_1, \dots, X_m$, for all $i \in \{0, 1, \dots, m\}$, $\mathrm{rad}(X_i) \le \frac{5}{8}\mathrm{rad}(X)$. By Corollary 4.5 we need to show that $\alpha \le 1/8$. Since $C = 8\sqrt{c \cdot \hat{c}}$, then

$$\alpha = \sqrt{\epsilon_{\lim}}/C = \frac{1}{8}\sqrt{\frac{|X|}{c|Y|}\left(\frac{\mathrm{rad}(Y)}{\mathrm{rad}(X)}\right)^{1/4}} \le \frac{1}{8}\sqrt{\left(\frac{\mathrm{rad}(X)}{\mathrm{rad}(Y)}\right)^{3/4}} \le \frac{1}{8}, \tag{4.11}$$

where the inequality holds since X is not a reset cluster.

In order to prove the second property, we can use the first property and obtain

$$\alpha \le \frac{1}{8}\left(\frac{\mathrm{rad}(X)}{\mathrm{rad}(Y)}\right)^{3/8} \le \frac{1}{8}\left(\frac{5}{8}\right)^{3t/8} \le \frac{1}{8}\left(\frac{7}{8}\right)^{t}. \qquad \square$$

We now show that given such a bound on α, which decreases exponentially with the number of levels from the last reset cluster, the spanning tree of each cluster increases its diameter by at most a constant factor. The main issue is that when there is a reset, the parameter α is "reset" to a constant, and so the total radius increase could potentially be very large. The key property which enables us to overcome this problem is that reset clusters have small radius. In particular, we will argue that for any cluster $X \in \mathcal{G}$, the sum of *all* the radii of reset clusters in $\mathcal{R}(X)$ is a constant factor smaller than $\mathrm{rad}(X)$.

We actually prove a more general statement, that the radius of the tree is bounded as long as the α parameters are a converging sequence, as this will be used later in the probabilistic embedding setting as well.

LEMMA 4.7. *If there exists $h : \mathbb{N} \to \mathbb{R}_+$ with $\sum_{t\ge 0} h(t) \le 1$, such that the hierarchical partition satisfies for all $X \in \mathcal{F}$, $\alpha = \alpha(X) \le h(L(X))$, then for any $F \in \mathcal{F}$, $\mathrm{rad}(T[F]) \le c' \cdot \mathrm{rad}(F)$.*

Proof. We first prove by induction on the construction tree $\mathcal{G}$ that for every $X \in \mathcal{G}$ with $t = L(X)$ (recall that this is the number of levels from the nearest ancestor reset cluster in the construction tree and is 0 for reset clusters),

$$\mathrm{rad}(T[X]) \le \mathrm{rad}(X) \cdot \prod_{j \ge t}(1 + h(j)) + \sum_{R \in \mathcal{R}(X)} \mathrm{rad}(T[R]). \tag{4.12}$$

The base case is when X is a leaf of $\mathcal{G}$; then the claim trivially holds as $\mathrm{rad}(T[X]) = 0$. Otherwise, we partition X into $X_0, \dots, X_m$ and assume by induction that the hypothesis is true for the children of X in $\mathcal{G}$. Let $i \in [m]$ be such that $d_X(y_i', x_i) + \mathrm{rad}(T[X_i])$ is maximal; then we claim that

$$\mathrm{rad}(T[X]) \le \mathrm{rad}(T[X_0]) + d_X(y_i', x_i) + \mathrm{rad}(T[X_i]). \tag{4.13}$$

To see this, assume $z \in X$ is the point such that $d_T(x_0, z) = \mathrm{rad}(T[X])$; then if $z \in X_j$ for some $j \ge 1$ we get that $d_T(x_0, z) = d_T(x_0, y_j') + d_T(y_j', x_j) + d_T(x_j, z) \le$

$\mathrm{rad}(T[X_0]) + d_X(y_j', x_j) + \mathrm{rad}(T[X_j]) \le \mathrm{rad}(T[X_0]) + d_X(y_i', x_i) + \mathrm{rad}(T[X_i])$ (the case $z \in X_0$ is trivial).

There are four cases to consider, whether X_0 and X_i are reset clusters or not. Consider first the case that both are not reset clusters. For $q \in \{0, i\}$, $L(X_q)$ is equal to $t + 1$, and so by the induction hypothesis,

$$\mathrm{rad}(T[X_q]) \le \mathrm{rad}(X_q) \cdot \prod_{j \ge t+1} (1 + h(j)) + \sum_{R \in \mathcal{R}(X_q)} \mathrm{rad}(T[R]). \tag{4.14}$$

Observe that if $R \in \mathcal{R}(X_q)$, then since X_q is not a reset cluster, $R \in \mathcal{R}(X)$ as well. Also, clearly $\mathcal{R}(X_0)$, $\mathcal{R}(X_i)$ are disjoint. Now, by Claim 4.4 we get that

$$\mathrm{rad}(X_0) + d_X(y_i', x_i) + \mathrm{rad}(X_i) \le \mathrm{rad}(X)(1 + \alpha) \le \mathrm{rad}(X)(1 + h(t)), \tag{4.15}$$

which yields

$$\begin{aligned} \mathrm{rad}(T[X]) &\overset{(4.13)}{\le} \mathrm{rad}(T[X_0]) + d_X(y_i', x_i) + \mathrm{rad}(T[X_i]) \\ &\overset{(4.14)}{\le} (\mathrm{rad}(X_0) + d_X(y_i', x_i) + \mathrm{rad}(X_i)) \prod_{j \ge t+1} (1 + h(j)) + \sum_{R \in \mathcal{R}(X_0) \cup \mathcal{R}(X_i)} \mathrm{rad}(T[R]) \\ &\overset{(4.15)}{\le} \mathrm{rad}(X)(1 + h(t)) \cdot \prod_{j \ge t+1} (1 + h(j)) + \sum_{R \in \mathcal{R}(X)} \mathrm{rad}(T[R]) \\ &= \mathrm{rad}(X) \cdot \prod_{j \ge t} (1 + h(j)) + \sum_{R \in \mathcal{R}(X)} \mathrm{rad}(T[R]). \end{aligned}$$

If X_i is a reset cluster and X_0 is not, then since $X_i \in \mathcal{R}(X) \setminus \mathcal{R}(X_0)$, a similar calculation gives that

$$\begin{aligned} \mathrm{rad}(T[X]) &\le (\mathrm{rad}(X_0) + d_X(y_i', x_i)) \prod_{j \ge t+1} (1 + h(j)) + \sum_{R \in \mathcal{R}(X_0)} (\mathrm{rad}(T[R])) + \mathrm{rad}(T[X_i]) \\ &\le \mathrm{rad}(X) \cdot \prod_{j \ge t} (1 + h(j)) + \sum_{R \in \mathcal{R}(X)} \mathrm{rad}(T[R]). \end{aligned}$$

The other cases, when X_i is a reset cluster and X_0 is not, and when both are reset clusters, are similar. This completes the proof of (4.12). Now we continue to prove the lemma. First, we prove by induction on the construction tree $\mathcal{G}$ that the lemma holds for the set of reset clusters. In fact we show a stronger bound, which is necessary in order to obtain the bound for nonreset clusters. Recall that $c = e + 1$ and $c' = 2e + 1$. We show that for every reset cluster $Y \in \mathcal{R}$ we have

$$\mathrm{rad}(T[Y]) \le c \cdot \mathrm{rad}(Y). \tag{4.16}$$

Assume the induction hypothesis is true for all descendants of Y in $\mathcal{R}$. In particular, for all $R \in \mathcal{R}(Y)$, $\mathrm{rad}(T[R]) \le c \cdot \mathrm{rad}(R)$. Recall that R becomes a reset cluster since $\mathrm{rad}(R) \le \frac{\mathrm{rad}(Y)}{c \cdot |Y|}|R|$, and using that $\{R : R \in \mathcal{R}(Y)\}$ are pairwise disjoint,[5]

$$\sum_{R \in \mathcal{R}(Y)} \mathrm{rad}(R) \le \frac{\mathrm{rad}(Y)}{c|Y|} \sum_{R \in \mathcal{R}(Y)} |R| \le \frac{\mathrm{rad}(Y)}{c}. \tag{4.17}$$

It follows that

$$\begin{aligned} \mathrm{rad}(T[Y]) &\overset{(4.12)}{\le} \mathrm{rad}(Y) \cdot \prod_{j \ge 0}(1 + h(j)) + \sum_{R \in \mathcal{R}(Y)} \mathrm{rad}(T[R]) \\ &\overset{(4.17) \wedge (4.16)}{\le} \mathrm{rad}(Y) \cdot e^{\sum_{j \ge 0} h(j)} + c \cdot \mathrm{rad}(Y)/c \\ &\le (e + 1)\mathrm{rad}(Y) = c \cdot \mathrm{rad}(Y). \end{aligned}$$

Finally, we show the lemma holds for all the other clusters. Let $F \in \mathcal{F} \setminus \mathcal{R}$ and $Y \in \mathcal{R}$ such that $F \in \mathcal{G}_Y$. Let $t = L(F)$. Note that $\sum_{R \in \mathcal{R}(F)} |R| \le |F|$. Since $F \notin \mathcal{R}$ we have $\frac{\mathrm{rad}(Y)}{c|Y|} \le \frac{\mathrm{rad}(F)}{|F|}$ and it follows that

$$\sum_{R \in \mathcal{R}(F)} \mathrm{rad}(R) \le \frac{\mathrm{rad}(Y)}{c|Y|} \sum_{R \in \mathcal{R}(F)} |R| \le \frac{\mathrm{rad}(Y)}{c|Y|} \cdot |F| \le \mathrm{rad}(F). \tag{4.18}$$

Finally,

$$\begin{aligned} \mathrm{rad}(T[F]) &\overset{(4.12)}{\le} \mathrm{rad}(F) \cdot \prod_{j \ge t}(1 + h(j)) + \sum_{R \in \mathcal{R}(F)} \mathrm{rad}(T[R]) \\ &\overset{(4.16)}{\le} e \cdot \mathrm{rad}(F) + c \sum_{R \in \mathcal{R}(F)} \mathrm{rad}(R) \\ &\overset{(4.18)}{\le} e \cdot \mathrm{rad}(F) + c \cdot \mathrm{rad}(F) \\ &= c' \cdot \mathrm{rad}(F), \end{aligned}$$

proving the lemma. $\square$

We now proceed to bound for every ϵ the number of pairs with distortion $\Omega(\sqrt{1/\epsilon})$, thus proving the scaling distortion of our constructed spanning tree. We begin with some definitions that will be crucial in the analysis.

Definition 4.8. *For each $\epsilon \in (0, 1)$ and $Y \in \mathcal{R}$ let $\mathcal{K}(Y, \epsilon) = \{F \in \mathcal{G}_Y \setminus \mathcal{R}(Y) : |F| < \epsilon/\hat{c} \cdot |Y|\}$.* In other words, a cluster is in $\mathcal{K}(Y, \epsilon)$ if it contains less than $\epsilon/\hat{c}$ fraction of the points of Y. The following proposition will be useful.

Proposition 4.9. *Fix a cluster F with reset ancestor Y. Then for any $\epsilon > \epsilon_{\lim}(F)$, $F \in \mathcal{K}(Y, \epsilon)$.*

Proof. This is immediate from the definition of $\epsilon_{\lim}$, $\epsilon > \epsilon_{\lim} = \frac{|F|}{\beta \cdot |Y|} \ge \frac{\hat{c} \cdot |F|}{|Y|}$. $\square$

Informally, when counting the badly distorted pairs for a given ϵ, whenever we reach a cluster in $\mathcal{K}(Y, \epsilon)$ we count all its pairs as bad. For $Y \in \mathcal{R}$ let $\mathcal{G}_{Y,\epsilon}$ be the subtree rooted at Y that contains all the nodes X whose path (in the construction

[5] Recall that the new Steiner nodes do not contribute to the cardinality of a set.

tree $\mathcal{G}$) to Y (excluding Y and X) contains no node in $\mathcal{R} \cup \mathcal{K}(Y, \epsilon)$. In other words, $\mathcal{G}_{Y,\epsilon}$ is the tree rooted at Y whose leaves are reset clusters and clusters in $\mathcal{K}(Y, \epsilon)$, such that all inner nodes (except for the root Y) are not reset clusters nor in $\mathcal{K}(Y, \epsilon)$. Observe that $\mathcal{G}_{Y,\epsilon}$ is a subtree of $\mathcal{G}_Y$.

Recall the definition of $B_\epsilon(X)$ in (4.1), where the term $\hat{B}_\epsilon(P_i)$ is defined as in Lemma 4.1, and it is the square of the number of points which are "close" to the partition in the cone distance. In the following lemma we bound $B_\epsilon(Y)$ for every reset cluster Y for any value of ϵ. Note that $B_\epsilon(Y)$ does not count all the distorted pairs, as there are some pairs which are distorted for values of $\epsilon \in [\bar{\epsilon}, \epsilon_{\lim}]$, and those will be accounted for by $\bar{B}_\epsilon(Y)$, which is bounded in Observation 4.11.

LEMMA 4.10. *For any $Y \in \mathcal{R}$, $\epsilon \in (0, 1)$ we have that $B_\epsilon(Y) \le \epsilon|Y|^2/4$.*

Proof. As mentioned above, we will argue that for any reset cluster Y, the number of pairs of points that are both contained in a cluster K for $K \in \mathcal{K}(Y, \epsilon)$ is sufficiently small so that we can ignore them all. The pairs that are contained in a reset cluster R for $R \in \mathcal{R}(Y)$ will be handled recursively. We need to handle pairs that may be distorted by some partition before reaching the leaves of $\mathcal{G}_Y$. To this end, we define for a cluster $X \in \mathcal{G}_Y$

$$\mathcal{E}_Y(X) = (X \times Y) \setminus \left(\bigcup_{R \in \mathcal{R}(Y) \cap \mathcal{G}_X} (R \times R) \right); \tag{4.19}$$

these are all the pairs, whose first point is in X, that were separated by the hierarchical star-partition algorithm before reaching the leaves of $\mathcal{G}_X$. Note that each pair in $X \times X$ is counted twice. We prove by induction on the construction tree $\mathcal{G}_{Y,\epsilon}$ that if $t = L_Y(X)$,

$$B_\epsilon(X) \le \epsilon/\hat{c} \cdot |\mathcal{E}_Y(X)| \sum_{j \ge t} (9/10)^j + \sum_{R \in \mathcal{R}(Y) \cap \mathcal{G}_X} B_\epsilon(R) + \sum_{K \in \mathcal{K}(Y,\epsilon) \cap \mathcal{G}_X} B_\epsilon(K). \tag{4.20}$$

The base of the induction, where X is a leaf in $\mathcal{G}_{Y,\epsilon}$, is trivially true, because in the base case either $X \in \mathcal{K}(Y, \epsilon)$ or $X \in \mathcal{R}(Y)$, and then as $X \in \mathcal{G}_X$ we have that $B_\epsilon(X)$ appears on the right-hand side of (4.20).

For the inductive step, assume that (4.20) holds for all the children $X_0, \dots, X_m$ of X. Let $P = \{P_i\}_{i=0}^{m-1}$ be the set of partitions that created these clusters, that is, $P_i = (X_i; Y_i)$, where $Y_i = X \setminus (\cup_{0 \le j \le i} X_j)$. Note that the value of $\epsilon_{\lim} = \epsilon_{\lim}(X)$ is the same for all the partitions P_i; however, the value of $\bar{\epsilon} = \bar{\epsilon}(i)$ returned by the decompose algorithm can be different for values of $0 \le i \le m - 1$. Since $X \notin \mathcal{K}(Y, \epsilon) \cup \mathcal{R}(Y)$, by Definition 4.8 we have that $\epsilon \le \hat{c} \cdot |X|/|Y| \le 1/\beta \cdot |X|/|Y| = \epsilon_{\lim}$. Hence we can apply Lemma 4.1 to deduce a bound on $\hat{B}_\epsilon(P_i)$. By Claim 4.6 we have $\beta = \frac{1}{\hat{c}}(\frac{\mathrm{rad}(X)}{\mathrm{rad}(Y)})^{1/4} \le \frac{1}{\hat{c}}(\frac{5}{8})^{t/4}$. From Lemma 4.1, and using that $\hat{B}_\epsilon(P_i) = 0$ for partitions P_i in which $\epsilon \ge \bar{\epsilon}(i)$, we obtain for every $0 \le i \le m - 1$ that

$$\hat{B}_\epsilon(P_i) \le \epsilon \cdot |X_i| \cdot |Y \setminus X_i| \cdot (5/8)^{t/4}/\hat{c}.$$

Observe that each pair in $X_i \times (Y \setminus X_i)$ cannot appear in $R \times R$ for any $R \in \mathcal{R}(Y) \cap \mathcal{G}_X$ because this pair is separated. Also, summing over all $0 \le i \le m - 1$ we count every pair of $X \times Y$ at most once, so that

$$\sum_{i=0}^{m-1} \hat{B}_\epsilon(P_i) \le \epsilon/\hat{c} \cdot (5/8)^{t/4} \sum_{i=0}^{m-1} |X_i| \cdot |Y \setminus X_i| \le \epsilon/\hat{c} \cdot (9/10)^t |\mathcal{E}_Y(X)|. \tag{4.21}$$

Since $X \notin \mathcal{K}(Y,\epsilon) \cup \mathcal{R}(Y)$ we have that each $R \in \mathcal{R}(Y) \cap \mathcal{G}_X$ (resp., $K \in \mathcal{K}(Y,\epsilon) \cap \mathcal{G}_X$) appears in exactly one of $\mathcal{R}(Y) \cap \mathcal{G}_{X_i}$ (resp., $\mathcal{K}(Y,\epsilon) \cap \mathcal{G}_{X_i}$) for some $0 \le i \le m$ (in other words, for each leaf of $\mathcal{G}_{Y,\epsilon}$ in the subtree rooted at X, there is exactly one X_i such that the leaf belongs to the subtree rooted at X_i). This implies the following:

$$\sum_{i=0}^{m} |\mathcal{E}_Y(X_i)| = |\mathcal{E}_Y(X)|, \tag{4.22}$$

$$\sum_{i=0}^{m} \left(\sum_{R \in \mathcal{R}(Y) \cap \mathcal{G}_{X_i}} B_\epsilon(R) \right) = \sum_{R \in \mathcal{R}(Y) \cap \mathcal{G}_X} B_\epsilon(R),$$

$$\sum_{i=0}^{m} \left(\sum_{K \in \mathcal{K}(Y,\epsilon) \cap \mathcal{G}_{X_i}} B_\epsilon(K) \right) = \sum_{K \in \mathcal{K}(Y,\epsilon) \cap \mathcal{G}_X} B_\epsilon(K).$$

The number of levels in the construction tree from each X_i to Y is $t+1$, and so

$$\begin{aligned}
B_\epsilon(X) &\overset{(4.1)}{=} \sum_{i=0}^{m-1} \hat{B}_\epsilon(P_i) + \sum_{i=0}^{m} B_\epsilon(X_i) \\
&\overset{(4.20)}{\le} \sum_{i=0}^{m-1} \hat{B}_\epsilon(P_i) + \sum_{i=0}^{m} \Bigg(\epsilon/\hat{c} \cdot |\mathcal{E}_Y(X_i)| \sum_{j \ge t+1} (9/10)^j \\
&\qquad + \sum_{R \in \mathcal{R}(Y) \cap \mathcal{G}_{X_i}} B_\epsilon(R) + \sum_{K \in \mathcal{K}(Y,\epsilon) \cap \mathcal{G}_{X_i}} B_\epsilon(K) \Bigg) \\
&\overset{(4.22)\wedge(4.21)}{\le} \epsilon/\hat{c} \cdot (9/10)^t |\mathcal{E}_Y(X)| + \epsilon/\hat{c} \cdot |\mathcal{E}_Y(X)| \sum_{j \ge t+1} (9/10)^j \\
&\qquad + \sum_{R \in \mathcal{R}(Y) \cap \mathcal{G}_X} B_\epsilon(R) + \sum_{K \in \mathcal{K}(Y,\epsilon) \cap \mathcal{G}_X} B_\epsilon(K) \\
&= \epsilon/\hat{c} \cdot |\mathcal{E}_Y(X)| \sum_{j \ge t} (9/10)^j + \sum_{R \in \mathcal{R}(Y) \cap \mathcal{G}_X} B_\epsilon(R) + \sum_{K \in \mathcal{K}(Y,\epsilon) \cap \mathcal{G}_X} B_\epsilon(K),
\end{aligned}$$

which proves the inductive claim.

We now prove the assertion of the lemma by induction on the construction tree $\mathcal{G}$. The base case for leaves of $\mathcal{G}$ is trivial, as they are of size 1 and contain no pairs. Let $Y \in \mathcal{R}$. By the induction hypothesis, for every $R \in \mathcal{R}(Y)$,

$$B_\epsilon(R) \le \epsilon |R|^2/4. \tag{4.23}$$

Observe that if $K \in \mathcal{K}(Y,\epsilon)$, then we treat all pairs in K as distorted, and using Definition 4.8

$$B_\epsilon(K) \le |K|^2 \le \frac{1}{\hat{c}} \cdot \epsilon |Y| \cdot |K|. \tag{4.24}$$

Since the clusters in $\mathcal{R}(Y) \cup \mathcal{K}(Y,\epsilon)$ are pairwise disjoint,

$$\sum_{R \in \mathcal{R}(Y)} |R| + \sum_{K \in \mathcal{K}(Y,\epsilon)} |K| \le |Y|. \tag{4.25}$$

Recall that $\hat{c} = 44$ and $\sum_{j\geq 0}(9/10)^j = 10$. Finally,

$$
\begin{aligned}
B_\epsilon(Y) &\overset{(4.20)}{\leq} \epsilon/\hat{c} \cdot |\mathcal{E}_Y(Y)| \sum_{j\geq 0}(9/10)^j + \sum_{R\in\mathcal{R}(Y)} B_\epsilon(R) + \sum_{K\in\mathcal{K}(Y,\epsilon)} B_\epsilon(K) \\
&\overset{(4.23)\wedge(4.24)}{\leq} 10\epsilon/44 \cdot |\mathcal{E}_Y(Y)| + \epsilon/4 \cdot \sum_{R\in\mathcal{R}(Y)} |R|^2 + \epsilon \cdot |Y|/44 \cdot \sum_{K\in\mathcal{K}(Y,\epsilon)} |K| \\
&= \left[10\epsilon/44 \cdot |\mathcal{E}_Y(Y)| + 10\epsilon/44 \sum_{R\in\mathcal{R}(Y)} |R|^2 \right] \\
&\quad + \left[\epsilon/44 \cdot \sum_{R\in\mathcal{R}(Y)} |R|^2 + \epsilon \cdot |Y|/44 \cdot \sum_{K\in\mathcal{K}(Y,\epsilon)} |K| \right] \\
&\overset{(4.19)}{\leq} 10\epsilon/44 \cdot |Y|^2 + \epsilon \cdot |Y|/44 \cdot \left(\sum_{R\in\mathcal{R}(Y)} |R| + \sum_{K\in\mathcal{K}(Y,\epsilon)} |K| \right) \\
&\overset{(4.25)}{\leq} 10\epsilon/44 \cdot |Y|^2 + \epsilon/44 \cdot |Y|^2 \\
&= \epsilon/4 \cdot |Y|^2. \qquad \square
\end{aligned}
$$

Observation 4.12. *For every $\epsilon \in (0, 1]$, $\bar{B}_\epsilon(G) \leq \epsilon n^2/8$.*

Proof. Recall the definition of $\bar{B}_\epsilon(G)$ in (4.2). By Lemma 4.1, for all $x \in V(G)$, $\bar{B}_\epsilon(x) \leq \epsilon n/8$, and thus

$$\bar{B}_\epsilon(G) = \sum_{x\in V(G)} \bar{B}_\epsilon(x) \leq \epsilon n^2/8. \qquad \square$$

Proof of Theorem 1.5. First we show that the total number of pairs whose distortion is too large is at most $\epsilon\binom{n}{2}$ for any value of $\epsilon \in (0, 1]$. Indeed, applying Lemma 4.10 on the original graph G suggests that $B_\epsilon(G) \leq \epsilon n^2/4$, and by Observation 4.11, $\bar{B}_\epsilon(G) \leq \epsilon n^2/8$.

For the distortion analysis, we have to be extremely careful: it could be that some $x, y \in X$ are not separated in the star-partition of X. However, in the cluster X_i that contains them, the induced distance is increased—this can happen if the shortest path between x, y in X is not fully contained in X_i. For this reason, we will argue that even if the shortest path between a pair is cut (that is, not fully contained in a single cluster), then either this pair is accounted for in $B_\epsilon(G)$ or $\bar{B}_\epsilon(G)$ or the following holds: Even if x, y will be in the maximal possible distance in the final tree, their distortion will be small enough.

Fix any $\epsilon \in (0, 1)$ and some pair $x, y \in V(G)$. Let $x = v_1, \ldots, v_t = y$ be a shortest path in G between x and y. Let X be a cluster for which this path is cut for the first time, that is, $v_1, \ldots, v_t \in X$, and when X is partitioned to $X_0, \ldots, X_m$, then there is some $0 \leq i \leq m$ such that $0 < |X_i \cap \{v_1, \ldots, v_t\}| < t$. We take the minimal such i, which means that $v_1, \ldots, v_t \in Y_{i-1}$. Let $\hat{\Lambda} = \mathrm{rad}(X)$. In order to create the cluster X_i, the `decompose` algorithm is called on Y_{i-1} with the cone metric $\rho = \ell_{x_i}^{x_0}$ (where the cone metric is with respect to the graph induced by $X_0 \cup Y_{i-1}$ and the set Y_{i-1}) and creates a partition $P = (X_i; Y_i)$ where $X_i = B_{(Y_{i-1},\rho)}(x_i, r)$ for some radius r. W.l.o.g. assume that $1 \leq j < t$ is such that $v_j \in X_i$, $v_{j+1} \notin X_i$ (the other possibility that $v_j \notin X_i$ and $v_{j+1} \in X_i$ is symmetric). If it is the case that $\epsilon > \epsilon_{\lim} = \epsilon_{\lim}(X)$, then by definition $B_\epsilon(X) = \binom{|X|}{2}$, and the pair $\{x, y\}$ is accounted for there. So from now on we assume that $\epsilon \le \epsilon_{\lim}$.

If it is the case that $d_G(x, y) < \sqrt{\epsilon}\hat{\Lambda}/(300C)$, then we will show that both $x, y \in S_\epsilon(P)$. To see this, consider first the case where $x \in X_i$, then using (4.4), $\rho(x, v_{j+1}) \le 2d_{Y_{i-1}}(x, v_{j+1}) = 2d_G(x, v_{j+1}) \le 2d_G(x, y) < \sqrt{\epsilon}\hat{\Lambda}/(150C)$, and since $\rho(x_i, v_{j+1}) > r$, by the triangle inequality $\rho(x_i, x) \ge \rho(x_i, v_{j+1}) - \rho(x, v_{j+1}) > r - \sqrt{\epsilon}\hat{\Lambda}/(150C)$ so we get that $x \in S_\epsilon(P)$. The other case is when $x \notin X_i$, then similarly $\rho(x, v_j) \le 2d_G(x, v_j) \le 2d_G(x, y) < \sqrt{\epsilon}\hat{\Lambda}/(150C)$, and as $\rho(x_i, v_j) \le r$ we obtain that $\rho(x_i, x) \le \rho(x_i, v_j) + \rho(v_j, x) < r + \sqrt{\epsilon}\hat{\Lambda}/(150C)$ so again $x \in S_\epsilon(P)$. The argument for y is analogous. Finally, we consider two cases. If $\epsilon \le \bar{\epsilon}$ (where $\bar{\epsilon}$ is the parameter returned by `decompose` when creating the partition P), then as $\hat{B}_\epsilon(P) = |S_\epsilon(P)|^2$, the pair x, y is accounted for in $\hat{B}_\epsilon(P)$. The other case is that $\bar{\epsilon} < \epsilon \le \epsilon_{\lim}$, then since $d_{Y_{i-1}}(x, y) < \sqrt{\epsilon}\hat{\Lambda}/(300C)$ we have that $y \in B_{Y_{i-1}}(x, \sqrt{\epsilon}\hat{\Lambda}/(150C))$. Note that if $\bar{B}_\epsilon(X)$ is defined in the current partition P, then y contributes to $\bar{B}_\epsilon(x)$ (in the sense that it appears in the appropriate ball in the definition of $\bar{B}_\epsilon(x)$), while if it has been defined in the previous iteration, while partitioning some cluster Y which is an ancestor of X, we claim that y already contributed to $\bar{B}_\epsilon(x)$. To see this, observe that $\bar{B}_\epsilon(x)$ depends only on ϵ, X and on $\hat{\Lambda}$, and by Corollary 4.5, the radius of Y is larger than $\hat{\Lambda}$. Since we are using the induced metric on X, the ball in Y as defined for $\bar{B}_\epsilon(x)$ will surely contain y as well.

We now argue that if $d_G(x, y) \ge \sqrt{\epsilon}\hat{\Lambda}/(300C)$ the distortion will be sufficiently small. This will follow once we establish that $d_T(x, y) \le 2c'\hat{\Lambda}$, in which case the distortion will be at most $600c'C/\sqrt{\epsilon} = O(1/\sqrt{\epsilon})$. To prove this, we use Lemma 4.7 with the parameter $h(t) = \frac{1}{8} \cdot (\frac{7}{8})^t$. This choice satisfies the conditions of the lemma since $\sum_{t \ge 0} h(t) = 1$, and using the second property of Claim 4.6 we have that indeed $\alpha(X) \le h(L(X))$. By the assertion of Lemma 4.7 we obtain $d_T(x, y) \le 2\mathrm{rad}(T[X]) \le 2c'\hat{\Lambda}$. Applying Lemma 2.2 yields the promised bounds on the ℓ_q distortion. ∎

Finally, we complete the proof of Lemma 4.1 stating the properties of our generic decompose algorithm.

Proof of Lemma 4.2. In what follows, unless stated explicitly, all the balls are with respect to the metric (W, ρ) (which may be a cone pseudometric). The proof is very similar to the proof of the ultrametric case, but there are two main differences. The first is that we need to satisfy the property of the lemma only for $\epsilon \le \epsilon_{\lim}$, so we can use the fact that $|W| \le \epsilon_{\lim} \cdot \beta \cdot n$ to obtain an improved bound on the number of (possibly) distorted pairs. The second difference is that we cannot choose the center point u; it is given as input. Recall that in the ultrametric case we chose a specific u so that $|B^\circ(u, \Delta/2)| \le n/2$ (in a cone metric, even a ball of radius 0 may contain an arbitrary number of points!); therefore we need to consider two cases. The first is that a ball of certain radius (analogous to the ball of half the radius in the ultrametric case) contains less than $n/2$ points, and we choose the radius in a similar manner to Claim 3.3, so that $|Z| \le n/2$. In the second case, the roles of Z and $\bar{Z} = W \setminus Z$ switch, and we choose the radius to be at least that certain radius, so that $|\bar{Z}| \le n/2$.

Note that the parameters $\bar{\epsilon}, \hat{\epsilon}$ defined below are not necessarily smaller than 1. Recall that $\alpha = \sqrt{\epsilon_{\lim}}/C$ and $\beta \le 1/\hat{c} = 1/44$.

Case 1. $|B(u, (\theta + \alpha/2)\hat{\Lambda})| \le n/2$. In this case let $\hat{\epsilon} = \max\{\epsilon \in (0, \epsilon_{\lim}] : |B(u, (\theta + \frac{\sqrt{\epsilon}}{4C})\hat{\Lambda})| \ge \epsilon \cdot \beta \cdot n\}$ (recall that such $\hat{\epsilon}$ exists, because when $\epsilon = 1/(\beta n)$ the condition is satisfied). Let $\hat{S} = [(\theta + \frac{\sqrt{\hat{\epsilon}}}{4C})\hat{\Lambda}, (\theta + \frac{\sqrt{\hat{\epsilon}}}{2C})\hat{\Lambda})$, and $S = [(\theta + \frac{\sqrt{\hat{\epsilon}}}{C}(\frac{1}{4} + \frac{1}{25}))\hat{\Lambda},$

$(\theta + \frac{\sqrt{\hat\epsilon}}{C}(\frac{1}{2} - \frac{1}{25}))\hat\Lambda]$. As in the previous section, we will choose $r \in S$ and define the partition by $Z = B(u,r)$. Observe that

$$\hat\epsilon \cdot \beta \cdot n \le |Z| \le n/2. \tag{4.26}$$

Case 2. $|B(u,(\theta+\alpha/2)\hat\Lambda)| > n/2$. In this case let $\hat\epsilon = \max\{\epsilon \in (0,\epsilon_{\lim}] \ : \ |W \setminus B(u,(\theta+\alpha-\frac{\sqrt{\epsilon}}{4C})\hat\Lambda)| \ge \epsilon\cdot\beta\cdot n\}$. Let $\hat S = [(\theta+\alpha-\frac{\sqrt{\hat\epsilon}}{2C})\hat\Lambda, (\theta+\alpha-\frac{\sqrt{\hat\epsilon}}{4C})\hat\Lambda]$, and $S = [(\theta+\alpha-\frac{\sqrt{\hat\epsilon}}{C}(\frac{1}{2}-\frac{1}{25}))\hat\Lambda, (\theta+\alpha-\frac{\sqrt{\hat\epsilon}}{C}(\frac{1}{4}+\frac{1}{25}))\hat\Lambda]$.

Note that choosing $r \in S$ and $Z = B(u,r)$ guarantees that

$$\hat\epsilon \cdot \beta \cdot n \le |\bar Z| \le n/2. \tag{4.27}$$

We show that one can choose $r \in S$ and define the partition $P = (Z;\bar Z)$ by $Z = B(u,r)$ such that the property of the lemma holds with $\bar\epsilon = 32\hat\epsilon$. The algorithm will return $\bar\epsilon$. First we show the property for $\epsilon \in [\bar\epsilon, \epsilon_{\lim}]$ for any $r \in S$ and either of the two cases. Let $x \in S_\epsilon(P)$ (as defined in the lemma).

Case 1. Note that since $\rho(u,x) \le r + \frac{\sqrt{\epsilon}\hat\Lambda}{150C}$ we have that $B(x, \frac{2\sqrt{\epsilon}\hat\Lambda}{150C}) \subseteq B(u, r + \frac{3\sqrt{\epsilon}\hat\Lambda}{150C}) \subseteq B(u, (\theta+\frac{\sqrt{\epsilon}}{4C})\hat\Lambda)$, and we used that $r < (\theta + \frac{\sqrt{\hat\epsilon}}{2C})\hat\Lambda \le (\theta + \frac{\sqrt{\epsilon/32}}{2C})\hat\Lambda$. By the maximality of $\hat\epsilon$ and since $\epsilon > \hat\epsilon$ we have that

$$\left|B\left(x, \frac{2\sqrt{\epsilon}\hat\Lambda}{150C}\right)\right| \le \left|B\left(u, \left(\theta + \frac{\sqrt{\epsilon}}{4C}\right)\hat\Lambda\right)\right| \le \epsilon\cdot\beta\cdot n < \epsilon n/8. \tag{4.28}$$

Case 2. This case is similar to the previous case. This time using that $\rho(u,x) \ge r - \frac{\sqrt{\epsilon}\hat\Lambda}{150C}$, we have that $B(x, \frac{2\sqrt{\epsilon}\hat\Lambda}{150C}) \subseteq W \setminus B^\circ(u, \rho(u,x) - \frac{2\sqrt{\epsilon}\hat\Lambda}{150C}) \subseteq W \setminus B^\circ(u, r - \frac{3\sqrt{\epsilon}\hat\Lambda}{150C}) \subseteq W \setminus B(u, (\theta+\alpha-\frac{\sqrt{\epsilon}}{4C})\hat\Lambda)$, where the last inequality is using that $r > (\theta+\alpha-\frac{\sqrt{\hat\epsilon}}{2C})\hat\Lambda \ge (\theta+\alpha-\frac{\sqrt{\epsilon/32}}{2C})\hat\Lambda$. By the maximality of $\hat\epsilon$ and since $\epsilon > \hat\epsilon$ it follows that

$$\left|B\left(x, \frac{2\sqrt{\epsilon}\hat\Lambda}{150C}\right)\right| \le \left|W \setminus B\left(u, \left(\theta + \alpha - \frac{\sqrt{\epsilon}}{4C}\right)\hat\Lambda\right)\right| \le \epsilon\cdot\beta\cdot n < \epsilon n/8. \tag{4.29}$$

Using (4.4) with (4.28) (or (4.29), depending on which case), we conclude that

$$\bar B_\epsilon(x) = \left|B_{(W,d_W)}\left(x, \frac{\sqrt{\epsilon}\hat\Lambda}{150C}\right)\right| \le \left|B\left(x, \frac{2\sqrt{\epsilon}\hat\Lambda}{150C}\right)\right| \le \epsilon n/8.$$

We next show that a certain choice of $r \in S$ will produce a partition that satisfies the property of the lemma for all $\epsilon \in (0,\bar\epsilon]$. The proof is very similar to that of Claim 3.3: we first bound the total number of points whose distance to u lies in S and then iteratively delete any "bad" interval in S, those that contain too many points. Then we argue that we must have run out of points before all of S could be removed. This suggests that any radius in the remaining interval will be good for all ϵ in the appropriate range. The proof here is slightly more involved, only because we have some extra parameters such as ϵ_{rmlim} and β that control the size of W, and we have two cases to consider, but the essential idea is the same.

For any $r \in S$ and $\epsilon \le \bar\epsilon$ let $S_r(\epsilon) = [r - \sqrt{\epsilon}\hat\Lambda/(150C)), r + \sqrt{\epsilon}\hat\Lambda/(150C))]$, $s(\epsilon) = \sqrt{\epsilon}\hat\Lambda/(75C)$, and let $Q_r(\epsilon) = \{w \in W \ : \ \rho(u,w) \in S_r(\epsilon)\}$. Note that the length of the interval S is given by $s = 17/(100C)\sqrt{\hat\epsilon}\hat\Lambda$ and that for any $r \in S$

and any $\epsilon \le \bar{\epsilon}$, $S_r(\epsilon) \subseteq \hat{S}$. We say that properly $A_r(\epsilon)$ holds if cutting at radius r is "good" for ϵ; formally: $A_r(\epsilon)$ iff $|Q_r(\epsilon)| \le \sqrt{\epsilon \cdot \hat{\epsilon}/2} \cdot n \cdot \beta$. As before we define $Q = \{w \in W \ : \ \rho(u,w) \in \hat{S}\}$.

PROPOSITION 4.12. *$|Q| \le 4\hat{\epsilon} \cdot \beta \cdot n$.*

Proof. In Case 1, we have that $Q \subseteq B(u, (\theta + \sqrt{\hat{\epsilon}}/(2C))\hat{\Lambda})$. We distinguish between two cases. If $\hat{\epsilon} \le \epsilon_{\lim}/4$, then $|B(u, (\theta + \sqrt{4\hat{\epsilon}}/(4C))\hat{\Lambda})| \le 4\hat{\epsilon} \cdot \beta \cdot n$ (by the maximality of $\hat{\epsilon}$). Otherwise, $\hat{\epsilon} > \epsilon_{\lim}/4$. In this case, by the restriction on $\epsilon_{\lim}$ imposed in the lemma, $|Q| \le |W| \le \epsilon_{\lim} \cdot \beta \cdot n \le 4\hat{\epsilon} \cdot \beta \cdot n$.

In Case 2, $Q \subseteq W \setminus B(u, (\theta + \alpha - \sqrt{\hat{\epsilon}}/(2C))\hat{\Lambda})$. We distinguish between two cases. If $\hat{\epsilon} \le \epsilon_{\lim}/4$, then $|W \setminus B(u, (\theta + \alpha - \sqrt{4\hat{\epsilon}}/(4C))\hat{\Lambda})| \le 4\hat{\epsilon} \cdot \beta \cdot n$ (by the maximality of $\hat{\epsilon}$). Otherwise, $\hat{\epsilon} > \epsilon_{\lim}/4$, and again $|Q| \le |W| \le 4\hat{\epsilon} \cdot \beta \cdot n$. $\square$

CLAIM 4.13. *There exists some $r \in S$ such that properly $A_r(\epsilon)$ holds for all $\epsilon \in (0, \bar{\epsilon}]$.*

Proof. The proof is very similar to the proof of Claim 3.3. We perform a process that deletes the "worst" interval from S. Initially, let $I_0 = S$, and $j = 1$.

1. If for all $r \in I_{j-1}$ and for all $\epsilon \in (0, \bar{\epsilon}]$ property $A_r(\epsilon)$ holds, then set $t = j - 1$, stop the iterative process, and output I_t.
2. Let $\mathcal{S}_j = \{S_r(\epsilon) \ : \ r \in I_{j-1}, \epsilon \le 32\hat{\epsilon}, \neg A_r(\epsilon)\}$. We greedily remove the interval that has maximal ϵ. Formally, let r_j, ϵ_j be parameters such that $S_{r_j}(\epsilon_j) \in \mathcal{S}_j$ and $\epsilon_j = \max\{\epsilon \ : \ \exists S_r(\epsilon) \in \mathcal{S}_j\}$.
3. Set $I_j = I_{j-1} \setminus S_{r_j}(\epsilon_j)$, set $j = j+1$, and goto 1.

We now argue that $I_t \ne \emptyset$. First observe that $\sum_{j=1}^{t} |Q_{r_j}(\epsilon_j)| \le 2|Q| \le 8\hat{\epsilon} \cdot \beta \cdot n$. Recall that since $A_{r_j}(\epsilon_j)$ does not hold, then for any $1 \le j \le t$: $|Q_{r_j}(\epsilon_j)| > \sqrt{\epsilon_j \cdot \hat{\epsilon}/2} \cdot \beta \cdot n$, which implies that $\sum_{j=1}^{t} \sqrt{\epsilon_j} < 12\sqrt{\hat{\epsilon}}$. Now we can bound the total length of the removed intervals,

$$\sum_{j=1}^{t} s(\epsilon_j) \le \sum_{j=1}^{t} \sqrt{\epsilon_j}\Delta/(75C) \le 12/(75C) \cdot \sqrt{\hat{\epsilon}}\Delta = 16/(100C) \cdot \sqrt{\hat{\epsilon}}\Delta.$$

Since $s = 17/(100C) \cdot \sqrt{\hat{\epsilon}}\Delta$, then indeed $I_t \ne \emptyset$, so any $r \in I_t$ satisfies the condition of the claim. $\square$

Claim 4.13 shows that for any $\epsilon \in (0, \bar{\epsilon}]$ we have

$$\hat{B}_\epsilon(P) \le \epsilon \cdot \hat{\epsilon}/2 \cdot (n \cdot \beta)^2 \le \epsilon \cdot \beta \cdot |Z| \cdot (n - |Z|);$$

the last inequality holds using (4.26) in Case 1, which also implies that $n/2 \le n - |Z|$. In Case 2 we are using (4.27), which yields $n/2 \le |Z|$ (and also that $|\bar{Z}| \le n - |Z|$). $\square$

5. Lower bound. In this section we show that our upper bounds on scaling distortion are tight even for the n-cycle.

LEMMA 5.1. *For any $\epsilon \in (1/n^2, 1)$, any $(1-\epsilon)$-partial embedding of the n-cycle into a tree requires distortion at least $\Omega(1/\sqrt{\epsilon})$.*

Proof. Fix some $\epsilon \in (1/n^2, 1)$, and let $v_1, \ldots, v_n$ be the (ordered) vertices of the n-cycle (X, d). The proof idea is to show that for any choice of $P \subseteq \binom{X}{2}$ with $|P| \ge (1-\epsilon)\binom{n}{2}$, we can find $k \approx 1/\sqrt{\epsilon}$ points on the cycle $u_1, \ldots, u_k$ such that the metric induced on $\{u_1, \ldots, u_k\}$ is almost a cycle metric, and all the pairs $\{u_i, u_j\}$ are in P. Applying the known lower bound of $\Omega(k)$ for embedding a k-cycle into any tree [RR98] will finish the proof.

Let $k = 1/(4\sqrt{\epsilon})$ and $m = n/k$.[6] Divide the vertices of the cycle into k consecutive parts $\bar{U}_1, \ldots, \bar{U}_k$, that is, for any $1 \le i \le k$ let $\bar{U}_i = \{v_{(i-1)m+1}, \ldots, v_{im}\}$. Let $U_i \subseteq \bar{U}_i$ be the central $m/2$ points of $\bar{U}_i$. Given any $P \subseteq \binom{X}{2}$, it is sufficient to find $u_i \in U_i$ for each $1 \le i \le k$ such that for any $1 \le i < j \le k$, $\{u_i, u_j\} \in P$ (because $d(u_i, u_j) \approx m \cdot \min\{j - i, i + k - j\}$ up to a factor of 4). We will choose these representatives u_i iteratively. At each step there will be a forbidden set of points $B_i \subseteq U_i$. Let $N(u)$ denote the set of points v such that $\{u, v\} \notin P$, and let $\deg(u) = |N(u)|$. We start with $i = 1$:

1. Let $B_i = U_i \cap (\bigcup_{j=1}^{i-1} N(u_j))$.
2. Choose $u_i \in U_i \setminus B_i$ with the minimum degree.
3. If $i < k$ let $i = i + 1$ and go to 1.

Now we show that for any i, $|B_i| \le m/4$, which will conclude the proof. Fix $1 \le i \le k$, and assume inductively for all $1 \le j < i$ that $|B_j| \le m/4$ (observe that $B_1 = \emptyset$). It is sufficient to show that $\sum_{j=1}^{i-1} \deg(u_j) \le m/4$, and in order to do so, consider the total number of pairs outside P, which must be at most $\epsilon n^2/2$. The minimality of $\deg(u_j)$ and the fact that $|U_j \setminus B_j| \ge m/4$ indicate that there are at least $m/4$ points in U_j of degree at least $\deg(u_j)$. Summing over all $1 \le j < i$, noticing that each pair may be counted twice, gives a total number of $m/8 \sum_{j=1}^{i-1} \deg(u_j)$ pairs outside P. Using that $\epsilon n^2/2 = m^2/32$ we get that

$$\sum_{j=1}^{i-1} \deg(u_j) \le 8/m \cdot \epsilon n^2/2 = m/4. \qquad \square$$

6. Probabilistic scaling embedding into spanning trees. In this section we prove Theorem 1.6. The proof of this theorem is based on a somewhat simpler variation of the decomposition algorithm from the previous section. In fact, the hierarchical star-partition algorithm remains practically the same, with modified submethod probabilistic star-partition given in Figure 3, instead of star-partition. Note that this method does not require n, the size of last reset cluster, as a parameter.

Let $f : [1, \infty) \to [1, \infty)$ be a monotone nondecreasing function satisfying $f(i) \ge i$ and

$$\sum_{i=1}^{\infty} \frac{1}{f(i)} \le 1/2. \tag{6.1}$$

For example, if we define $\log^{(0)} n = n$, and for any $i > 0$ define recursively $\log^{(i)} n = \max\{\log(\log^{(i-1)} n), 1\}$, then we can take for any constants $\theta > 0$, $t \in \mathbb{N}$ the function $f(x) = \hat{c} \prod_{j=0}^{t-1} \log^{(j)}(x) \cdot (\log^{(t)}(x))^{1+\theta}$, for sufficiently large constant $\hat{c} > 0$, and it will satisfy the conditions.

6.1. Algorithm analysis. Let $\hat{\mathcal{H}}$ be the distribution on laminar families induced by the algorithm above. Let $\mathcal{H} = \mathrm{supp}(\hat{\mathcal{H}})$. We begin with a bound on the radius of any tree that can be created by the randomized algorithm. In what follows we fix a certain family $\mathcal{F} \in \mathcal{H}$ with a corresponding construction tree $\mathcal{G}$ and use the definitions of the first paragraph in section 4.1. We say that a node $X \in \mathcal{F}$ is in *level* i of $\mathcal{G}$ if the distance in the construction tree from X to the root is i. We have the following claims analogous to Claim 4.6.

[6] Assume w.l.o.g. that k is an integer and m is an even integer.

$(X_0, \ldots, X_t, (y_1, x_1), \ldots, (y_t, x_t)) = \texttt{probabilistic-star-partition}(X, x_0, \Lambda)$:

1. Set $k = 0$; $\hat{\Lambda} = \mathrm{rad}_{x_0}(X)$; $\alpha = \frac{1}{16 f(\log(2\Lambda/\hat{\Lambda}))}$;
2. Choose uniformly at random $\beta \in [1/2, 5/8]$.
3. $X_0 = B(x_0, \beta\hat{\Lambda})$; $Y_0 = X \setminus X_0$;
4. If $Y_k = \emptyset$ set $t = k$ and stop; Otherwise, set $k = k + 1$;
5. Let $v_k \in Y_{k-1}$ be the point minimizing $\hat{\chi}_k = \frac{|X|}{|B_{Y_{k-1}}(x, \alpha\hat{\Lambda}/2)|}$ over all $x \in Y_{k-1}$; Set $\chi_k = \max\{4, \hat{\chi}_k\}$;
6. Choose $r_k \in [\alpha\hat{\Lambda}/2, \alpha\hat{\Lambda}]$ according to the following random process:
   - Divide the interval $[\alpha\hat{\Lambda}/2, \alpha\hat{\Lambda}]$ into $N = \lceil 2 \log \chi_k \rceil$ equal length intervals $J_1, \ldots, J_N$; Let $h = 1$;
   - LOOP: Toss a fair coin; If it turns out head and $h < N$ then let $h = h + 1$ and goto LOOP;
   - Choose r_k uniformly at random from the interval J_h.
7. Let $\{x_k, y_k\}$ be the edge in E which lies on a shortest path from v_k to x_0 such that $y_k \in X_0, x_k \in Y_{k-1}$ [a];
8. Let $\ell = \ell_{x_k}^{x_0}$ be the cone-metric with respect to x_0 and x_k on the subspace Y_{k-1};
   $X_k = B_{(Y_{k-1}, \ell)}(x_k, r_k)$; $Y_k = Y_{k-1} \setminus X_k$.
9. goto 4;

[a] By Claim 4.3, if $z_k \in Y_{k-1}$ all the points on any shortest path from v_i to x_0 are either in X_0 or in Y_{k-1}

FIG. 3. *Probabilistic star-partition algorithm.*

CLAIM 6.1. *Fix some $Y \in \mathcal{R}$. Let $X \in \mathcal{G}_Y \setminus \mathcal{R}(Y)$ with $L_Y(X) = k$, and let $Z \in \mathcal{G}_X$ with $L_Y(Z) = k + l$. Then*

- $\mathrm{rad}(Z) \leq (\frac{5}{8})^l \mathrm{rad}(X)$,
- $\alpha = \alpha(X) \leq \frac{1}{16 f(1+k/2)}$.

Proof. The proof of the first property is essentially identical to the proof of the first property of Claim 4.6. We prove by induction on l that the base case $l = 0$ holds as then $Z = X$. Assume for $l - 1$ and for l that using Corollary 4.5 it suffices to show that $\alpha \leq 1/8$. This indeed holds because $\Lambda \geq \hat{\Lambda}$ and $f(i) \geq i$.

In order to prove the second property we use the first property and obtain

$$\alpha = \frac{1}{16 f(\log(2\mathrm{rad}(Y)/\mathrm{rad}(X)))} \leq \frac{1}{16 f(1 + k \log\left(\frac{8}{5}\right))} \leq \frac{1}{16 f(1 + k/2)}. \qquad \square$$

Observe that the function $h : \mathbb{N} \to \mathbb{R}_+$ defined by $h(i) = \frac{1}{8 f(1+i/2)}$ satisfies both conditions of Lemma 4.7, that is, for any $X \in \mathcal{F}$,

- $\sum_{i \geq 0} h(i) = \sum_{i \geq 0} \frac{1}{8 f(1+i/2)} \leq \frac{1}{4} \sum_{i \geq 1} \frac{1}{f(i)} \leq 1$ and
- $\alpha(X) \leq h(L(X))$.

We conclude that for any $X \in \mathcal{F}$,

$$\mathrm{rad}[T(X)] \leq c' \cdot \mathrm{rad}(X). \tag{6.2}$$

Having established a bound on the radius increase of any tree in the support of the distribution, we continue to bound the expected distortion of $(1 - \epsilon)$ fraction of the pairs for all $0 < \epsilon \le 1$ simultaneously. Recall that $G = (V, E)$ is the original graph we work with. Unless stated explicitly otherwise, all distances are with respect to the shortest path metric on G. Let $\Lambda = \mathrm{rad}(G)$. For any $\epsilon > 0$ and $x \in X$ define $r_\epsilon(x)$ as the minimal radius r for which $|B(x,r)| \ge \epsilon n$. Fix some $\epsilon > 0$, and let $P(\epsilon) = \{\{x,y\} : d(x,y) \ge \max\{r_{\epsilon/2}(x), r_{\epsilon/2}(y)\}\}$; as stated in Definition 1.2, these are the pairs for which we want to bound the expected distortion. Observe that for any point x there are at most $\epsilon n/2$ other points y for which $d(x,y) < r_{\epsilon/2}(x)$ (and that the $\epsilon n/2$ closest distances are counted twice), so that $P(\epsilon) \ge \binom{n}{2} - (\epsilon n^2/2 - \epsilon n/2) = (1-\epsilon)\binom{n}{2}$. Throughout the analysis, a pair $\{x,y\} \in P(\epsilon)$ is fixed, and let $B = B(x, d(x,y))$. By definition of $P(\epsilon)$ we have that $|B| \ge \epsilon n/2$. For $i \in \mathbb{N}$ and $X \subseteq V$, let $S_{X,i} = S_{X,i}(B)$ be the event that X is in level i (i.e., a node of depth i in the construction tree) with $B \subseteq X$.

As before, a cluster X is partitioned into the central ball X_0 and cones $X_1, \dots X_m$, where m is a random variable depending on X and the random partitioning of X. For an integer j, let $\mathcal{E}_j(X,i)$ be the event that $S_{X,i}$ holds, $B \cap X_j \notin \{\emptyset, B\}$ and for all $k < j$, $B \cap X_k = \emptyset$. In other words, this is the first time in the hierarchical partition that the ball B is cut. Let $\mathcal{E}(X,i)$ be the event that $\exists\, 0 \le j < m$ such that $\mathcal{E}_j(X,i)$.

Remark. The ball $B(x, d(x,y))$ is taken according to the metric induced by G. Since we required that $B \subseteq X$, the distance between x, y remains the same in the subgraph induced on X as it was in G.

Fix some cluster X with $B \subseteq X$. For a subset $Y \subseteq X$ and integer $j \ge 0$ let $R_{Y,j} = R_{Y,j}(X,i)$ be the event that $S_{X,i}$ holds, and in the star-partition of the subgraph X, it happens that $Y = Y_{j-1}$ and $B \subseteq Y$. Let $\mathcal{C}_{Y,j} = \mathcal{C}_{Y,j}(X,i)$ be the event that $R_{Y,j}(X,i)$ holds and in addition $B \cap X_j \neq \emptyset$. In other words, this is the first time that the ball B is either cut or fully contained in a cluster when performing the partitioning of X. Note that there are always unique Y and j such that this event holds. Let $\mathcal{T}$ be the support of the distribution over spanning trees induced by the algorithm. Note that the events $\mathcal{E}(X,i)$ are disjoint for different X or i and that they form a partition of the (implicit) probability space (because the ball B has $|B| \ge 2$, so it must be cut for the first time exactly once), so we can write

$$\begin{aligned}
\mathbb{E}[d_T(x,y)] &= \sum_{T\in\mathcal{T}} \Pr[T]\cdot d_T(x,y) \qquad (6.3)\\
&= \sum_{T\in\mathcal{T}}\sum_{i\in\mathbb{N}}\sum_{X\subseteq V} \Pr[\mathcal{E}(X,i)]\cdot \Pr[T \mid \mathcal{E}(X,i)]\cdot d_T(x,y)\\
&= \sum_{i\in\mathbb{N}}\sum_{X\subseteq V} \Pr[\mathcal{E}(X,i)] \sum_{T\in\mathcal{T}} \Pr[T \mid \mathcal{E}(X,i)]\cdot d_T(x,y)\\
&\le 2c'\sum_{i\in\mathbb{N}}\sum_{X\subseteq V} \Pr[\mathcal{E}(X,i)]\cdot \mathrm{rad}(X) \sum_{T\in\mathcal{T}} \Pr[T \mid \mathcal{E}(X,i)]\\
&= 2c'\sum_{i\in\mathbb{N}}\sum_{X\subseteq V} \Pr[\mathcal{E}(X,i)]\cdot \mathrm{rad}(X). \qquad (6.4)
\end{aligned}$$

The inequality holds since conditioning on $\mathcal{E}(X,i)$ suggests that $B \subseteq X$, which means that both $x, y \in X$ so that $d_T(x,y) = d_{T[X]}(x,y) \le 2\mathrm{rad}(T[X])$, and by (6.2) it follows that $\mathrm{rad}(T[X]) \le c'\mathrm{rad}(X)$. The main technical lemma is the following.

LEMMA 6.2. *There exists a universal constant C' such that for any cluster X and integer i,*

$$\Pr[\mathcal{E}(X,i)] \cdot \mathrm{rad}(X) \le C' \cdot f(\log(2/\epsilon)) d(x,y) \times \sum_{Y \subseteq X} \sum_{j \ge 0} \Pr[\mathcal{C}_{Y,j}(X,i)] \log\left(\frac{|X|}{|B_Y(x, \alpha \cdot \mathrm{rad}(X)/2)|}\right), \tag{6.5}$$

where $\alpha = \alpha(X)$ is defined as in the algorithm.

Before proving this lemma, let us show how it implies Theorem 1.6.

Proof of Theorem 1.6. Let $k = \lceil \log_{8/5}(128 f(\log(2c/\epsilon))) \rceil$. We will divide the summation in the last line of (6.3) into k summations, that is, for $\ell \in \{0, 1, \ldots, k-1\}$ the ℓth sum will be over indices $i \in I_\ell$, where $I_\ell = \{i \ : \ i = \ell \pmod k\}$. Fix such an ℓ, and we will prove by (reverse) induction on $t \in I_\ell$ that

$$\sum_{i \in I_\ell,\ i \ge t} \sum_X \Pr[\mathcal{E}(X,i)] \mathrm{rad}(X) \le 2C' \cdot f(\log(2/\epsilon)) d(x,y) \sum_X \Pr[S_{X,t}] \log\left(\frac{|X|}{\epsilon n/2}\right). \tag{6.6}$$

The base case will be when $t = \log_{8/5} \Lambda$. The first property of Claim 6.1 suggests that all the clusters in level t are singletons, so none of them can contain B, and thus $\Pr[\mathcal{E}(X,t)] = 0$ for all X. Assume (6.6) holds for $t+k$, and we prove for t. In the summation, we will distinguish between clusters X with $\mathrm{rad}(X) < 2d(x,y)/\alpha$ and the others. For the former clusters, it holds that $\sum_{X \ : \ \mathrm{rad}(X) < 2d(x,y)/\alpha} \Pr[\mathcal{E}(X,t)] \mathrm{rad}(X) \le 2d(x,y)/\alpha$, and for the latter we shall use Lemma 6.2. Using the induction hypothesis,

$$\begin{aligned}
&\sum_{i \in I_\ell,\ i \ge t} \sum_X \Pr[\mathcal{E}(X,i)] \mathrm{rad}(X) \\
&= \sum_X \Pr[\mathcal{E}(X,t)] \mathrm{rad}(X) + \sum_{i \in I_\ell,\ i \ge t+k} \sum_Z \Pr[\mathcal{E}(Z,i)] \mathrm{rad}(Z) \\
&\overset{(6.6)}{\le} \sum_{X \ : \ \mathrm{rad}(X) \ge 2d(x,y)/\alpha} \Pr[\mathcal{E}(X,t)] \mathrm{rad}(X) + \sum_{X \ : \ \mathrm{rad}(X) < 2d(x,y)/\alpha} \Pr[\mathcal{E}(X,t)] \mathrm{rad}(X) \\
&\quad + 2C' \cdot f(\log(2/\epsilon)) d(x,y) \sum_Z \Pr[S_{Z,t+k}] \log\left(\frac{|Z|}{\epsilon n/2}\right) \\
&\overset{(6.5)}{\le} \sum_{X \ : \ \mathrm{rad}(X) \ge 2d(x,y)/\alpha} C' \cdot f(\log(2/\epsilon)) d(x,y) \\
&\quad \times \sum_{Y \subseteq X} \sum_{j \ge 0} \Pr[\mathcal{C}_{Y,j}(X,t)] \log\left(\frac{|X|}{|B_Y(x, \alpha \cdot \mathrm{rad}(X)/2)|}\right) + 2d(x,y)/\alpha \\
&\quad + 2C' \cdot f(\log(2/\epsilon)) d(x,y) \sum_Z \Pr[S_{Z,t+k}] \log\left(\frac{|Z|}{\epsilon n/2}\right) \\
&= C' \cdot f(\log(2/\epsilon)) d(x,y) \\
&\quad \cdot \left(\sum_{X \ : \ \mathrm{rad}(X) \ge 2d(x,y)/\alpha} \sum_{Y \subseteq X} \sum_{j \ge 0} \Pr[\mathcal{C}_{Y,j}(X,t)] \log\left(\frac{|X|}{|B_Y(x, \alpha \cdot \mathrm{rad}(X)/2)|}\right) \right. \\
&\quad \left. + \sum_Z \Pr[S_{Z,t+k}] \log\left(\frac{|Z|}{\epsilon n/2}\right) \right) \\
&\quad + 2d(x,y)/\alpha + C' \cdot f(\log(2/\epsilon)) d(x,y) \sum_Z \Pr[S_{Z,t+k}] \log\left(\frac{|Z|}{\epsilon n/2}\right).
\end{aligned}$$

Thus, to prove the induction hypothesis it remains to show that

$$2/\alpha \leq C' \cdot f(\log(2/\epsilon))d(x,y) \tag{6.7}$$

and that

$$\sum_{X:\mathrm{rad}(X)\geq 2d(x,y)/\alpha} \sum_{Y\subseteq X} \sum_{j\geq 0} \Pr[\mathcal{C}_{Y,j}(X,t)] \log\left(\frac{|X|}{|B_Y(x,\alpha\cdot \mathrm{rad}(X)/2)|}\right) \\ + \sum_Z \Pr[S_{Z,t+k}] \log\left(\frac{|Z|}{\epsilon n/2}\right) \leq \sum_X \Pr[S_{X,t}] \log\left(\frac{|X|}{\epsilon n/2}\right). \tag{6.8}$$

Observe that any cluster Z in level $t+k$ for which $S_{Z,t+k}$ holds has an ancestor X at level t and a $Y \subseteq X$ and integer $j \geq 0$ such that $\mathcal{C}_{Y,j}(X,t)$ holds. The ancestor profile of Z, $a(Z) = (X,Y,j)$, is so if the cluster Z at level $t+k$ has such ancestors at level t. We will break the summation over all Z according to the ancestor profile. Observe that $\Pr[\mathcal{C}_{Y,j}(X,t)] \geq \sum_{Z:a(Z)=(X,Y,j)} \Pr[S_{Z,t+k}]$, because the events $S_{Z,t+k}$ are disjoint for different Z (the ball B must be contained in Z), and in fact the inequality is strict if the ball B can be cut in one of the levels $t, t+1, \ldots, t+k-1$. Let $q = q(X,Y,j) = \sum_{Z:a(Z)=(X,Y,j)} \Pr[S_{Z,t+k}]$. The reason for choosing k levels between X and Z is that using Claim 6.1, we have that

$$\mathrm{rad}(Z) \leq (5/8)^k \mathrm{rad}(X) = \mathrm{rad}(X)/(128 f(\log(2c/\epsilon))). \tag{6.9}$$

Now $\alpha = \frac{1}{16 f(\log(2\mathrm{rad}(F)/\mathrm{rad}(X)))}$, where F is the last reset cluster before X. By definition of reset cluster, $\mathrm{rad}(F)/\mathrm{rad}(X) \leq cn/|X|$, and since $\epsilon n/2 \leq |B| \leq |X|$ we have that $\mathrm{rad}(F)/\mathrm{rad}(X) \leq 2c/\epsilon$. This suggests that

$$\alpha \geq \frac{1}{32 f(\log(4c/\epsilon))},$$

which proves (6.7) for C' large enough. We also have that if $Z \subseteq Y$ and $S_{Z,t+k}$ holds, then $x \in Z$ and thus using (6.9),

$$Z = B_Z(x, 2\mathrm{rad}(Z)) \subseteq B_Y(x, \alpha \cdot \mathrm{rad}(X)/2).$$

We conclude that

$$\begin{aligned}\sum_Z \Pr[S_{Z,t+k}] \log\left(\frac{|Z|}{\epsilon n/2}\right) &= \sum_{X,Y,j} \sum_{Z:a(Z)=(X,Y,j)} \Pr[S_{Z,t+k}] \log\left(\frac{|Z|}{\epsilon n/2}\right) \\ &\leq \sum_{X,Y,j} \sum_{Z:a(Z)=(X,Y,j)} \Pr[S_{Z,t+k}] \log\left(\frac{|B_Y(x,\alpha\cdot\mathrm{rad}(X)/2)|}{\epsilon n/2}\right) \\ &= \sum_{X,Y,j} q(X,Y,j) \cdot \log\left(\frac{|B_Y(x,\alpha\cdot\mathrm{rad}(X)/2)|}{\epsilon n/2}\right).\end{aligned}$$

Plugging this into (6.8) and noting that since $\mathrm{rad}(X) \geq 2d(x,y)/\alpha$ and $B \subseteq Y$ we have that $|B_Y(x, \alpha \cdot \mathrm{rad}(X)/2)| \geq |B| \geq \epsilon n/2$. Now,

$$\begin{aligned}
&\sum_{X:\mathrm{rad}(X)\geq 2d(x,y)/\alpha}\sum_{Y\subseteq X}\sum_{j\geq 0}\Pr[\mathcal{C}_{Y,j}(X,t)]\log\left(\frac{|X|}{|B_Y(x,\alpha\cdot\mathrm{rad}(X)/2)|}\right)\\
&\qquad+\sum_{Z}\Pr[S_{Z,t+k}]\log\left(\frac{|Z|}{\epsilon n/2}\right)\\
&\leq \sum_{X:\mathrm{rad}(X)\geq 2d(x,y)/\alpha}\sum_{Y,j}\big(\Pr[\mathcal{C}_{Y,j}(X,t)]-q(X,Y,j)\big)\cdot\log\left(\frac{|X|}{\epsilon n/2}\right)\\
&\quad+\sum_{X,Y,j}q(X,Y,j)\cdot\left(\log\left(\frac{|X|}{|B_Y(x,\alpha\cdot\mathrm{rad}(X)/2)|}\right)+\log\left(\frac{|B_Y(x,\alpha\cdot\mathrm{rad}(X)/2)|}{\epsilon n/2}\right)\right)\\
&=\sum_{X,Y,j}\Pr[\mathcal{C}_{Y,j}(X,t)]\cdot\log\left(\frac{|X|}{\epsilon n/2}\right)\\
&=\sum_{X}\Pr[S_{X,t}]\cdot\log\left(\frac{|X|}{\epsilon n/2}\right)\sum_{Y,j}\Pr[\mathcal{C}_{Y,j}(X,t)\mid S_{X,t}]\\
&=\sum_{X}\Pr[S_{X,t}]\cdot\log\left(\frac{|X|}{\epsilon n/2}\right),
\end{aligned}$$

where in the last equality we used that events $\mathcal{C}_{Y,j}(X,t)$ are disjoint, and fixing X at level t their sum is 1 (the ball B must be cut or contained for some Y and j). This concludes the proof of (6.6), and we can finally show a bound on the expected distortion,

$$\begin{aligned}
\mathbb{E}[d_T(x,y)] &\overset{(6.3)}{\leq} 2c'\sum_{i\in\mathbb{N}}\sum_{X}\Pr[\mathcal{E}(X,i)]\cdot\mathrm{rad}(X)\\
&= 2c'\sum_{\ell=0}^{k-1}\sum_{i\in I_\ell}\sum_{X}\Pr[\mathcal{E}(X,i)]\cdot\mathrm{rad}(X)\\
&\overset{(6.6)}{\leq} 4c'C'\cdot f(\log(2/\epsilon))d(x,y)\sum_{\ell=0}^{k-1}\sum_{X}\Pr[S_{X,\ell}]\log\left(\frac{|X|}{\epsilon n/2}\right)\\
&\leq 4c'C'\cdot f(\log(2/\epsilon))d(x,y)\log(2/\epsilon)\sum_{\ell=0}^{k-1}\sum_{X}\Pr[S_{X,\ell}].
\end{aligned}$$

Since $S_{X,\ell}$ are disjoint events when ℓ is fixed, their sum over all X can be at most 1, so we conclude that

$$\mathbb{E}[d_T(x,y)] = O(f(\log(2/\epsilon))\cdot\log(2/\epsilon)\cdot k\cdot d(x,y)) = \tilde{O}(\log^2(2/\epsilon)\cdot d(x,y)). \qquad \square$$

It remains to prove Lemma 6.2. Fix some cluster X with radius $\hat{\Lambda}$ and integer i such that $S_{X,i}$ holds. We partition X into $X_0, X_1, \ldots, X_m$. Divide the event $\mathcal{E}(X,i)$ into three events:

- The first is the event that B is cut by the first cluster (the central ball X_0). This event is denoted by $\mathcal{B}$.
- The second is the event $\mathcal{L}$, that there exists $j>0$ and $Y\subseteq X$ such that $R_{Y,j}$ holds, $B\cap X_j\notin\{\emptyset, B\}$, and $r_j\in J_N$. In other words, it is the event that B is indeed cut in level i, by a cone whose radius is chosen from the last interval (see line 6 in Figure 3).

- The third event $\mathcal{M}$ is the completion of the first two events, that there exists $j > 0$ and $Y \subseteq X$ such that $R_{Y,j}$ holds, $B \cap X_j \notin \{\emptyset, B\}$, and $r_j \notin J_N$. In other words, it is the event that B is cut in level i, by a cone whose radius is not chosen from the last interval.

CLAIM 6.3. $\Pr[\mathcal{B}] \le O(d(x,y)/\hat{\Lambda})$.

Proof. The radius of the central ball is chosen uniformly at random from an interval of length $\hat{\Lambda}/8$. Let $a, b \in B$ be the closest and farthest points from x_0; then by the triangle inequality $d_X(x_0,b) - d_X(x_0,a) \le d_X(a,b) \le 2d(x,y)$. The probability that $\beta\hat{\Lambda}$ falls in an interval of length $2d(x,y)$ is at most $16d(x,y)/\hat{\Lambda}$. $\square$

CLAIM 6.4. $\Pr[\mathcal{L}] \le O(d(x,y)/(\alpha\hat{\Lambda}))$.

Proof. Fix some $j > 0$, and consider the interval J_N as defined in the algorithm. It has length $\alpha\hat{\Lambda}/(2N)$, so conditioning on the event that the radius is chosen from the last interval J_N, the probability that the ball B is cut, is bounded by $2d(x,y)/(\alpha\hat{\Lambda}/(2N)) = 4N \cdot d(x,y)/(\alpha\hat{\Lambda})$. The probability that the last interval is chosen is the probability that $N = \lceil 2\log\chi_j(Y)\rceil$ random fair coins came up heads, which is at most $1/\chi_j(Y)^2$, where $\chi_j = \chi_j(Y)$ is as defined in line 5 of Figure 3, and $Y = Y_{j-1}$. Now, using that $\chi_j(Y) \ge 4$, we can write $N \le 3\log\chi_j(Y)$, so that

$$\begin{aligned}
\Pr[\mathcal{L}] &= \sum_{Y\subseteq X}\sum_{j>0}\Pr[\mathcal{R}_{Y,j}(X,i) \wedge r_j \in J_N \wedge B \cap X_j \notin \{\emptyset, B\}] \\
&= \sum_{Y\subseteq X}\sum_{j>0}\Pr[\mathcal{R}_{Y,j}]\cdot\Pr[r_j \in J_N \mid \mathcal{R}_{Y,j}]\cdot\Pr[B\cap X_j \notin\{\emptyset,B\} \mid \mathcal{R}_{Y,j}\wedge r_j \in J_N] \\
&\le \sum_{Y\subseteq X}\sum_{j>0}\Pr[\mathcal{R}_{Y,j}]\frac{1}{\chi_j(Y)^2}\cdot\frac{12\log\chi_j(Y)\cdot d(x,y)}{\alpha\hat{\Lambda}} \\
&\le \frac{12d(x,y)}{\alpha\hat{\Lambda}}\sum_{Y\subseteq X}\sum_{j>0}\Pr[\mathcal{R}_{Y,j}]\frac{1}{\chi_j(Y)}.
\end{aligned}$$

It remains to show that $\sum_Y\sum_{j>0}\Pr[\mathcal{R}_{Y,j}]\frac{1}{\chi_j(Y)} \le 1$. For each possible $m \in \mathbb{N}$ and a partition of X to $X_0, X_1, \ldots, X_m$, we can write its probability in terms of the $\mathcal{R}_{Y,j}$ in the following manner. For every $m > 0$ and a sequence $X_0, X_1, \ldots, X_m$, let $\Pr[X_0, X_1, \ldots, X_m]$ be the probability that this is the partition of X (conditioning on the cluster X). Recall that $\chi_j(Y) \ge \hat{\chi}_j(Y) = \frac{|X|}{|B_{Y_{j-1}}(v_j, \alpha\hat{\Lambda}/2)|}$ (see the definition in Figure 1), and as $r_j \ge \alpha\hat{\Lambda}/2$ we have that $B_{Y_{j-1}}(v_j, \alpha\hat{\Lambda}/2) \subseteq X_j$ for any choice of r_j. This implies that

$$\chi_j(Y) \ge \frac{|X|}{|X_j|}. \tag{6.10}$$

Next, note that the probability of a certain event $Y = Y_{j-1}$ is equal to the sum of probabilities of sequences $X_0, \ldots, X_m$, over all sequences for which $Y = (X_j \cup \cdots \cup X_m)$, which means that

$$\begin{aligned}
\sum_Y\sum_{j>0}\Pr[\mathcal{R}_{Y,j}]\frac{1}{\chi_j(Y)} &\le \sum_Y\sum_{j>0}\Pr[Y = Y_{j-1}]\frac{1}{\chi_j(Y)} \\
&= \sum_Y\sum_{j>0}\sum_{m\ge j}\sum_{(X_0,\ldots,X_m)\,:\,Y=(X_j\cup\cdots\cup X_m)}\Pr[X_0,\ldots,X_m]\frac{1}{\chi_j(Y)} \\
&\overset{(6.10)}{\le} \sum_Y\sum_{j>0}\sum_{m\ge j}\sum_{(X_0,\ldots,X_m)\,:\,Y=(X_j\cup\cdots\cup X_m)}\Pr[X_0,\ldots,X_m]\frac{|X_j|}{|X|}.
\end{aligned} \tag{6.11}$$

Observe that any sequence $X_0, \ldots, X_m$ appears exactly m times in the summations, since for every $j = 1, \ldots, m$ there is a unique choice of Y such that $Y = (X_j \cup \cdots \cup X_m)$. We conclude that (6.11) is equal to

$$\begin{aligned}
&\sum_{j>0} \sum_{m \geq j} \sum_{(X_0,\ldots,X_m)} \Pr[X_0, \ldots, X_m] \frac{|X_j|}{|X|} \\
&\quad = \sum_{m>0} \sum_{(X_0,\ldots,X_m)} \Pr[X_0, \ldots, X_m] \sum_{j=1}^{m} \frac{|X_j|}{|X|} \\
&\quad \leq \sum_{m>0} \sum_{(X_0,\ldots,X_m)} \Pr[X_0, \ldots, X_m] \\
&\quad = 1.
\end{aligned}$$

The inequality is using that the $\{X_j\}$ are pairwise disjoint, and the last equality is that the events of obtaining each sequence $X_0, \ldots, X_m$ are disjoint. ∎

CLAIM 6.5. $\Pr[\mathcal{M}] \leq O(d(x,y)/(\alpha\hat{\Lambda})) \sum_{Y \subseteq X} \sum_{j \geq 0} \Pr[\mathcal{C}_{Y,j}(X,i)] \log(\frac{|X|}{|B_Y(x,\alpha\hat{\Lambda}/2)|})$.

Proof. Naturally $\Pr[r_j \notin J_N] \leq 1$, even conditioned on anything. Fix some Y and j with $B \subseteq Y_{j-1} = Y$ such that $B \cap X_j \neq \emptyset$ (that is, event $\mathcal{C}_{Y,j}$ holds); then we argue that the probability that $X_j \cap B \neq B$ (that is, the ball B is cut by the next cluster) is bounded by $\frac{16 \log \chi_j(Y) \cdot d(x,y)}{\alpha\hat{\Lambda}}$. Indeed, if the radius r_j is chosen from the interval J_h (whose length is $\alpha\hat{\Lambda}/(2N)$), as B is a ball of radius $d(x,y)$, the probability that a uniform choice of r_j in this interval will cut B is at most $2d(x,y)/|J_h| \leq \frac{16 \log \chi_j(Y) \cdot d(x,y)}{\alpha\hat{\Lambda}}$.

$$\begin{aligned}
\Pr[\mathcal{M}] &\leq \sum_{Y \subseteq X} \sum_{j>0} \Pr[\mathcal{R}_{Y,j}(X,i) \wedge B \cap X_j \notin \{\emptyset, B\}] \\
&= \sum_{Y \subseteq X} \sum_{j>0} \Pr[\mathcal{R}_{Y,j} \wedge B \cap X_j \neq \emptyset] \cdot \Pr[B \cap X_j \neq B \mid \mathcal{R}_{Y,j} \wedge B \cap X_j \neq \emptyset] \\
&= \sum_{Y \subseteq X} \sum_{j>0} \Pr[\mathcal{C}_{Y,j}] \cdot \Pr[B \cap X_j \neq B \mid \mathcal{C}_{Y,j}] \\
&\leq \sum_{Y \subseteq X} \sum_{j>0} \Pr[\mathcal{C}_{Y,j}] \cdot \frac{16 \log \chi_j(Y) \cdot d(x,y)}{\alpha\hat{\Lambda}}.
\end{aligned}$$

The proof is complete recalling that v_j was the point minimizing $\chi_j(Y)$, thus $\chi_j(Y) = \frac{|X|}{|B_Y(v_j,\alpha\hat{\Lambda}/2)|} \leq \frac{|X|}{|B_Y(x,\alpha\hat{\Lambda}/2)|}$. ∎

Proof of Lemma 6.2. The proof is done simply by noting that event $\mathcal{E}(X,i)$ is equal to the union of events $\mathcal{B}$, $\mathcal{L}$, and $\mathcal{M}$ and applying Claims 6.3, 6.4, and 6.5. ∎

REFERENCES

[ABC+05] I. ABRAHAM, Y. BARTAL, H. T.-H. CHAN, K. DHAMDHERE, A. GUPTA, J. M. KLEINBERG, O NEIMAN, AND A SLIVKINS, *Metric embeddings WITH relaxed guarantees*, in Proceedings of FOCS, IEEE Computer Society, 2005, pp. 83–100.

[ABN06] I. ABRAHAM, Y. BARTAL, AND O. NEIMAN, *Advances in metric embedding theory*, in Proceedings of the 38th Annual ACM Symposium on Theory of Computing, ACM Press, New York, 2006, pp. 271–286.

[ABN07] 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 Annual ACM-SIAM Symposium on Discrete Algorithms, 2007, pp. 502–511.

[ABN08] I. Abraham, Y. Bartal, and O. Neiman, Nearly tight low stretch spanning trees, in Proceedings of the 49th Annual IEEE Symposium on the Foundations of Computer Science, 2008, pp. 781–790.

[AKPW95] N. Alon, R. M. Karp, D Peleg, and D West, *A graph-theoretic game and its application to the k-server problem*, SIAM J. Comput., 24 (1995), pp. 78–100.

[AN12] I. Abraham and O. Neiman, Using petal-decompositions to build a low stretch spanning tree, in Proceedings of the 44th Symposium on Theory of Computing, ACM, New York, 2012, pp. 395–406.

[AS03] V. Athitsos and S. Sclaroff, *Database indexing methods for 3d hand pose estimation*, in Proceedings of the Gesture Workshop, 2003, pp. 288–299.

[Bar96] Y. Bartal, *Probabilistic approximation of metric spaces and its algorithmic applications*, in Proceedings of the 37th Annual Symposium on the Foundations of Computer Science, Burlington, VT, 1996, pp. 184–193.

[Bar98] Y. Bartal, *On approximating arbitrary metrics by tree metrics*, in Proceedings of the 30th Annual ACM Symposium on Theory of Computing, 1998, pp. 183–193.

[Bar04] Y. Bartal, *Graph decomposition lemmas and their role in metric embedding methods*, in 12th Annual European Symposium on Algorithms, 2004, pp. 89–97.

[BBM06] Y. Bartal, B. Bollobás, and M. Mendel, *Ramsey-type theorems for metric spaces with applications to online problems*, J. Comput. System Sci., 72 (2006), pp. 890–921.

[BLMN05] Y. Bartal, N. Linial, M. Mendel, and A. Naor, *On metric ramsey-type phenomena*, Ann. of Math, 162 (2005), pp. 643–709.

[BM02] Y. Bartal and Mendel, *On Low Dimensional Lipschitz Embeddings of Ultrametrics*, manuscript, 2002.

[BM03] Y. Bartal and M. Mendel, *Multi-embedding and path approximation of metric spaces*, in Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms, 2003, pp. 424–433.

[DPK82] N. Deo, G. Prabhu, and M. S. Krishnamoorthy, *Algorithms for generating fundamental cycles in a graph*, ACM Trans. Math. Software, 8 (1982), pp. 26–42.

[EEST05] M. Elkin, Y. Emek, D. A. Spielman, and S.-H. Teng, *Lower-stretch spanning trees*, in Proceedings of the 37th Annual ACM Symposium on Theory of Computing, 2005, pp. 494–503.

[ELR07] M. Elkin, C. Liebchen, and R. Rizzi, *New length bounds for cycle bases*, Inform. Process. Lett., 104 (2007), pp. 186–193.

[FRT03] J. Fakcharoenphol, S. Rao, and K. Talwar, *A tight bound on approximating arbitrary metrics by tree metrics*, in Proceedings of the 35th Annual ACM Symposium on Theory of Computing, 2003, pp. 448–455.

[HBK+03] E. Halperin, J. Buhler, R. M. Karp, R. Krauthgamer, and B. Westover, *Detecting protein sequence conservation via metric embeddings*, Bioinformatics, 19 (2003), pp. 122–129.

[HFC00] G. Hristescu and M. Farach-Colton, *COFE: A scalable method for feature extraction from complex objects*, in Proceedings of DaWaK, 2000, pp. 358–371.

[HPM06] S. Har-Peled and M. Mendel, *Fast construction of nets in low-dimensional metrics and their applications*, SIAM J. Comput, 35 (2006), pp. 1148–1184.

[HS00] G. Hjaltason and H. Samet, *Contractive Embedding Methods for Similarity Searching in Metric Spaces*, University of Maryland, College Park, 2000.

[Ind01] P. Indyk, *Algorithmic applications of low-distortion geometric embeddings*, in Proceedings of the 42nd Annual Symposium on Foundations of Computer Science, 2001, pp. 10–33.

[Kru64] J. B. Kruskal, *Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis*, Psychometrika, 29 (1964), pp. 1–27.

[KSW04] J. M. Kleinberg, A. Slivkins, and T. Wexler, *Triangulation and embedding using small sets of beacons*, in Proceedings of FOCS, 2004, pp. 444–453.

[KW78] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, SAGE Publications, Thousand Oaks, CA, 1978.

[RR98] Y. Rabinovich and R. Raz, *Lower bounds on the distortion of embedding finite metric spaces in graphs*, Discrete Comput. Geom., 19 (1998), pp. 79–94.

[ST04] Y. Shavitt and T. Tankel, *Big-bang simulation for embedding network distances in euclidean space*, IEEE/ACM Trans. Networking, 12 (2004), pp. 993–1006.

[TC04] L. Tang and M. Crovella, Geometric exploration of the landmark selection problem, in Proceedings of PAM, 2004, pp. 63–72.

[WLB+98] B. Y. Wu, G. Lancia, V. Bafna, K.-M. Chao, R. Ravi, and C. Y. Tang, *A polynomial time approximation scheme for minimum routing cost spanning trees*, in Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, 1998, pp. 21–32.