

MIT Open Access Articles

Euclidean Spanners in High Dimensions

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Har-Peled, Sariel, Piotr Indyk, and Anastasios Sidiropoulos. "Euclidean Spanners in High Dimensions." Khanna, Sanjeev, ed. Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2013.

As Published: <http://dx.doi.org/10.1137/1.9781611973105.57>

Publisher: Society for Industrial and Applied Mathematics

Persistent URL: <http://hdl.handle.net/1721.1/87001>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.



Euclidean Spanners in High Dimensions

Sariel Har-Peled*

Piotr Indyk†

Anastasios Sidiropoulos‡

Abstract

A classical result in metric geometry asserts that any n -point metric admits a linear-size spanner of dilation $O(\log n)$ [PS89]. More generally, for any $c > 1$, any metric space admits a spanner of size $O(n^{1+1/c})$, and dilation at most c . This bound is tight assuming the well-known girth conjecture of Erdős [Erd63].

We show that for a metric induced by a set of n points in high-dimensional Euclidean space, it is possible to obtain improved dilation/size trade-offs. More specifically, we show that any n -point Euclidean metric admits a near-linear size spanner of dilation $O(\sqrt{\log n})$. Using the LSH scheme of Andoni and Indyk [AI06] we further show that for any $c > 1$, there exist spanners of size roughly $O(n^{1+1/c^2})$ and dilation $O(c)$. Finally, we also exhibit super-linear lower bounds on the size of spanners with constant dilation.

1 Introduction

Given a metric $M = (X, \rho)$, a graph $G = (X, E)$ is a c -spanner for M if every for pair $p, q \in X$ the shortest path distance ρ_G in the graph G approximates the original distance $\rho(p, q)$ up to a factor of c , i.e., $\rho(p, q) \leq \rho_G(p, q) \leq c \cdot \rho(p, q)$. Of particular interest are spanners that are *sparse*, i.e., that contain a sub-quadratic (ideally linear) number of edges. Spanners are natural and useful representations of a metric, and as such they have been a subject of extensive research (see surveys [Epp00, Zwi01]). In particular, it is known that any metric admits a $(2k-1)$ -spanner of size $O(n^{1+1/k})$ for any integer $k > 0$ [PS89]; assuming the girth conjecture of Erdős [Erd63], this bound is tight. For simpler metrics that are induced by a set of n points in a low-dimensional Euclidean space (say, of dimension

d), the distortion bound can be improved considerably: there exists a $(1+\epsilon)$ -spanner with only $O(n(1/\epsilon)^{O(d)})$ edges [Sal91, Vai91]. For a constant dimension d , this gives a bound that is linear in n . In particular, there is quite a bit of work on spanners in low dimensional Euclidean space (see the book by Narasimhan and Smid [NS07]). Extension of these techniques implies spanners for metric with low doubling dimension [HM06].

In this paper we focus on a class of metrics that lie in between the above two extrema. Specifically, we consider metrics induced by a set of n points in the Euclidean space with *unbounded* dimension, and ask what is the best sparsity of a t -spanner achievable for such metrics. Perhaps surprisingly, to the best of our knowledge, no non-trivial results for such metrics were known.

Our results. Our first result is a construction of an $O(c)$ -spanner with $O(n^{1+1/c^2} \log^2 n)$ edges. For large values of c , this improves over the $n^{1+O(1/c)}$ bound for general metric spaces. In particular, this shows that any Euclidean metric admits a $O(\sqrt{\log n})$ -spanner of near-linear size. We also give a (simpler) construction that achieves sparsity $O(n^{1+1/c^2} \log n \log \Delta)$, where Δ is the *spread* of the metric M , i.e., the ratio between the largest and the smallest non-zero distances in M . The latter spanners use only two hops, i.e., for any two points $p, q \in X$, there is a path in G of length $\leq c \cdot \rho(p, q)$ that uses only two edges.

We complement the result by showing that any 2-hop c -spanner for such metrics must use at least $n^{1+\Omega(1/c^2)}$ edges. Thus, the exponent in our sparsity upper bound is asymptotically tight, at least for 2-hop spanners.

Our techniques. Our upper bound uses *locality-sensitive hashing* for the Euclidean metric due to Andoni and Indyk [AI06]; for $c = \sqrt{\log n}$, one can alternatively use the Lipschitz partitions of Charikar *et al.* [CCG⁺98]. Both techniques are used to cover the input point set with clusters of a specified diameter $\delta \cdot c$, such that any pair of points within a distance δ from each other are included in one of the clusters. By constructing such partitions for $O(\log \Delta)$ values of δ and connecting them into a graph, the bound follows.

Our lower bound proceeds by analyzing the isoperimetric properties of the Euclidean space. In particu-

*Department of Computer Science; University of Illinois; 201 N. Goodwin Avenue; Urbana, IL, 61801, USA; sariel@uiuc.edu; <http://www.uiuc.edu/~sariel/>. Work on this paper was partially supported by a NSF AF awards CCF-0915984 and CCF-1217462.

†Computer Science and Artificial Intelligence Lab; Massachusetts Institute of Technology; 32 Vassar Street; Cambridge, MA, 02139, USA; indyk@theory.csail.mit.edu.

‡Department of Computer Science; University of Illinois; 201 N. Goodwin Avenue; Urbana, IL, 61801, USA; sidiropo@illinois.edu.

lar, we use the tools from the work of Motwani *et al.* [MNP07], which showed that the parameters of the locality-sensitive hash functions of Andoni and Indyk [AI06] are asymptotically tight. Note that we do not prove that sparse spanners imply good LSH families, as such a reduction is not likely to exist. E.g., metrics induced by bounded-degree expander graphs have sparse spanners by definition, but it can be easily seen that good LSH families for such metrics do not exist.

2 Near-linear spanners with dilation $O(\sqrt{\log n})$

The basic idea behind our construction is to randomly partition the point-set into clusters of low diameter, connect every cluster into a star, and do it repeatedly to guarantee that all points in certain resolution are connected. Repeating this in all relevant resolutions results in the desired spanner.

DEFINITION 2.1. (LIPSCHITZ PARTITION) Let (X, ρ) be a metric space, and let F be a distribution over partitions of X . We say that F is (β, δ) -Lipschitz if the following conditions are satisfied:

- (i) For any partition $P \in \text{supp}(F)$, for any cluster $C \in P$, we have $\text{diam}(C) \leq \delta$.
- (ii) For every $x, y \in X$,

$$\Pr_{P \in F} [P(x) \neq P(y)] \leq \beta \frac{\rho(x, y)}{\delta},$$

where for every $z \in X$, $P(z)$ denotes the cluster of P containing x .

LEMMA 2.1. ([CCG⁺98]) For every $\delta > 0$, and $d \geq 1$, there exists a $(O(\sqrt{d}), \delta)$ -Lipschitz partition of $(\mathbb{R}^d, \|\cdot\|_2)$.

LEMMA 2.2. (JL LEMMA [JL84]) For any $\varepsilon > 0$, every set of n points in Euclidean space admits an embedding into $(\mathbb{R}^{O(\log n / \varepsilon^2)}, \|\cdot\|_2)$, with distortion $1 + \varepsilon$.

2.1 2-Hop spanners We first show how to obtain a 2-hop spanner, with density depending on the spread of the input metric. The **spread** of a point set X , is the ratio between the diameter of X and the closest pair distance of X .

THEOREM 2.1. Let X be a set of n points in Euclidean space, with spread Δ . Then, there exists a 2-hop spanner for $(X, \|\cdot\|_2)$ with dilation $O(\sqrt{\log n})$, and with $O(n \log n \log \Delta)$ edges.

Proof. By Lemma 2.2 we may assume, up to a constant loss in the final dilation, that $X \subset \mathbb{R}^d$, where $d =$

$c_1 \log n$, for some fixed $c_1 > 0$. We can further assume w.l.o.g. that the minimum distance between points in X is 1.

By Lemma 2.1, we have that there exists some constant $c_2 > 0$, such that for every $\delta > 0$, there exists a $(c_2 \sqrt{c_1 \log n}, \delta)$ -Lipschitz partition of $(X, \|\cdot\|_2)$. For every $i = \{0, \dots, \log \Delta\}$, fix a $(c_2 \sqrt{c_1 \log n}, 2^{i+1} c_2 \sqrt{c_1 \log n})$ -Lipschitz partition F_i of $(X, \|\cdot\|_2)$.

We can now construct a spanner $G = (X, E)$ as follows. We begin with a graph containing no edges. For every $i \in \{0, \dots, \log \Delta\}$ we sample from F_i , k partitions independently, $P_{i,1}, \dots, P_{i,k}$ of X , for some $k = O(\log n)$. For every $P_{i,j}$ in the resulting family of partitions, for every cluster $C \in P_{i,j}$, we pick a vertex $r(C) \in C$ and we connect $r(C)$ to every vertex $z \in C$ with an edge of length $\|r(C) - z\|_2$. This concludes the construction of G .

We now analyze the resulting dilation. Fix a pair of points $x, y \in X$. Suppose $\|x - y\|_2 = h \in [2^i, 2^{i+1})$. Observe that

$$\Pr_{P \in F_i} [P(x) \neq P(y)] \leq 1/2.$$

Since we sample $k = O(\log n)$ partitions from F_i , it follows that with high probability there exists some $j \in \{1, \dots, k\}$, such that $P_{i,j}(x) = P_{i,j}(y)$. Conditioned on this event, we have

$$\begin{aligned} \rho_G(x, y) &\leq O\left(\sqrt{\log n} \cdot \text{diam}(P_{i,j}(x))\right) \\ &\leq O\left(\sqrt{\log n} \cdot \|x - y\|_2\right). \end{aligned}$$

By taking a union bound over all pairs of points we see that G has dilation $O(\sqrt{\log n})$ with positive probability, which concludes the proof.

The bound on the number of edges of G follows, since for every random partition $P_{i,j}$, we add a forest in G . Each such forest has at most $n - 1$ edges, and there are $O(\log n \log \Delta)$ partitions in total.

2.2 Removing the dependence on the spread

We now show how to remove the dependence on the spread, by increasing the number of hops. We recall the following standard definition.

DEFINITION 2.2. (NET) Let (X, ρ) be a metric space, and let $\delta > 0$. A maximal set $Y \subseteq X$, such that for any $x, y \in Y$, $\rho(x, y) > \delta$, is called a **δ -net** for (X, ρ) .

THEOREM 2.2. Let X be a set of n points in Euclidean space. Then, there exists a spanner for $(X, \|\cdot\|_2)$ with dilation $O(\sqrt{\log n})$, and with $O(n \log n \log \log n)$ edges.

Proof. By Lemma 2.2 we may assume, up to a constant loss in the final dilation, that $X \subset \mathbb{R}^d$, where $d = c_1 \log n$, for some fixed $c_1 > 0$. We may further assume w.l.o.g. that the minimum distance between pairs in X is 1, and the diameter of X is Δ .

By Lemma 2.1, we have that there exists some constant $c_2 > 0$, such that for every $\delta > 0$, there exists a $(c_2\sqrt{c_1 \log n}, \delta)$ -Lipschitz partition of $(X, \|\cdot\|_2)$. For every $i = \{0, \dots, \lceil \log \Delta \rceil\}$, fix a $(c_2\sqrt{c_1 \log n}, 2^{i+1}c_2\sqrt{c_1 \log n})$ -Lipschitz partition F_i of $(X, \|\cdot\|_2)$.

We can now construct a spanner $G = (X, E)$ as follows. We begin with a graph containing no edges. For every $i \in \{0, \dots, \lceil \log \Delta \rceil + 2\}$, let $N_i \subseteq X$ be a 2^{i-2} -net for $(X, \|\cdot\|_2)$. Since a net is defined to be a maximal subset of X satisfying a certain property, we can pick each N_i so that

$$X = N_0 \supseteq N_1 \supseteq \dots \supseteq N_{\lceil \log \Delta \rceil + 2}.$$

We sample from F_i , and independently, k partitions $P_{i,1}, \dots, P_{i,k}$ of N_i , for some $k = O(\log n)$. For every $P_{i,j}$ in the resulting family of partitions, for every cluster $C \in P_{i,j}$, we pick a vertex $r(C) \in C$ and we connect $r(C)$ to every vertex $z \in C$ with an edge of length $\|r(C) - z\|_2$. This concludes the construction of G .

We now analyze the resulting dilation. Let $\alpha = 2^5 c_2 \sqrt{c_1}$. We prove that the dilation is at most $\alpha\sqrt{\log n}$ for all pairs $x, y \in X$, by performing induction on $\|x - y\|_2$. Fix a pair of points $x, y \in X$. By the inductive hypothesis, for any $z, w \in X$, with $\|z - w\|_2 < \|x - y\|_2$, we have $\|z - w\|_2 \leq \alpha\sqrt{\log n}$. Suppose that $\|x - y\|_2 \in [2^i, 2^{i+1})$. Pick $x', y' \in N_i$, such that $\|x - x'\|_2 \leq 2^{i-2}$, and $\|y - y'\|_2 \leq 2^{i-2}$. Observe that

$$\|x - x'\|_2 \leq 2^{i-2} < 2^i \leq \|x - y\|_2$$

Therefore, by the induction hypothesis, we have

$$(2.1) \quad \rho_G(x, x') \leq \alpha\sqrt{\log n}\|x - x'\|_2$$

Similarly, we obtain that

$$(2.2) \quad \rho_G(y, y') \leq \alpha\sqrt{\log n}\|y - y'\|_2$$

By the triangle inequality, we have

$$\begin{aligned} \|x' - y'\|_2 &\leq \|x - y\|_2 + \|x - x'\|_2 + \|y - y'\|_2 \\ &< 2^{i+1} + 2^{i-2} + 2^{i-2} < 2^{i+2} \end{aligned}$$

It follows that

$$\Pr_{P \in F_{i+2}} [P(x') \neq P(y')] < 1/2.$$

Since we sample $k = O(\log n)$ partitions from F_{i+2} , it follows that with high probability there exists some

$t \in \{1, \dots, k\}$, such that $P_{i+2,t}(x') = P_{i+2,t}(y')$. Let $C = P_{i+2,t}(x') = P_{i+2,t}(y')$ be the cluster containing both x' , and y' . Conditioned on this event, we have

$$\begin{aligned} \rho_G(x', y') &\leq \rho_G(x', r(C)) + \rho_G(y', r(C)) \\ &\leq 2\text{diam}(C) \\ (2.3) \quad &\leq 2^{i+4}c_2\sqrt{c_1 \log n} \end{aligned}$$

Combining (2.1), (2.2), and (2.3), we obtain

$$\begin{aligned} \rho_G(x, y) &\leq \rho_G(x', y') + \rho_G(x, x') + \rho_G(y, y') \\ &\leq 2^{i+4}c_2\sqrt{c_1 \log n} + \alpha\sqrt{\log n}\|x - x'\|_2 \\ &\quad + \alpha\sqrt{\log n}\|y - y'\|_2 \\ &< 2^{i+4}c_2\sqrt{c_1 \log n} + \alpha\sqrt{\log n}2^{i-1} \\ &\leq \alpha\sqrt{\log n}\|x - y\|_2. \end{aligned}$$

By taking a union bound over all pairs of points we see that G has dilation at most $\alpha\sqrt{\log n} = O(\sqrt{\log n})$ with positive probability.

It remains to bound the number of edges in G . Let $\delta = c_2\sqrt{c_1 \log n}$. Let $h = \lceil \log \Delta \rceil + 2$, and let $I = \{0, \dots, h\}$. We define an auxiliary tree T with $V(T) = \bigcup_{i \in I} (\{i\} \times N_i)$. For every $i \in I \setminus \{0\}$, for every $x \in N_i$, we have an edge in T between (i, x) , and $(i-1, x)$. For every $i \in I \setminus \{0\}$, for every $y \in N_{i-1} \setminus N_i$, there exists $y' \in N_i$, with $\|y - y'\|_2 \leq 2^{i-2}$. We add an edge in T between $(i-1, y)$, and (i, y') . We consider T as being rooted at the vertex (h, x^*) , where $N_h = \{x^*\}$. Consider a branch $B = (s, x), (s-1, x), \dots, (s-l, x)$ in T , such that all vertices in B have exactly one child in T , with $l > \log \delta$. This means that for any $i \in \{s - \log \delta - 2, \dots, s - l\}$, the ball of radius $2^{i+2}\delta$ around x in N_i , contains only x . Recall that F_i is supported on partitions of N_i into clusters of diameter at most $2^{i+1}\delta$. Therefore, for every partition $P_{i,j}$ chosen by the algorithm, we have $P_{i,j}(x) = \{x\}$, i.e. the cluster containing x does not contain any other points. Let U be the set of vertices $v \in V(T)$ such that either v has at least two children, or there exists an ancestor $u \in V(T)$ of v with at least two children, such that the distance between u and v in T is at most $\log \delta$. We have $|U| = O(n \log \delta)$. When we pick a random partition $P_{i,j} \in F_i$, the total number of edges added to G due to $P_{i,j}$ is $\sum_{C \in P_{i,j}} |C| - 1$. This quantity is upper-bounded by the total number of points that are contained in non-singleton clusters. By the above discussion, all such points are contained in U . Therefore, the total number of edges added in G , that correspond to a partition $P_{i,j}$, is $O(|U \cap (\{i\} \times N_i)|)$. Since we sample $O(\log n)$ partitions at every level i , it follows that the total number of edges added to G due to partitions at level i , is $O(|U \cap (\{i\} \times N_i)| \cdot \log n)$. Summing over all levels

$i \in I$, we obtain

$$\begin{aligned} |E(G)| &\leq \sum_{i \in I} O(|U \cap (\{i\} \times N_i)| \cdot \log n) \\ &= O(|U| \cdot \log n) \\ &= O(n \log n \log \delta) \\ &= O(n \log n \log \log n), \end{aligned}$$

concluding the proof.

3 General dilation

Our construction for the general dilation uses the notion of *locality-sensitive hashing* [HIM12].

DEFINITION 3.1. (LOCALITY-SENSITIVE HASHING) *Let H be a family of hash functions mapping \mathbb{R}^d to some universe U . We say that H is $(\delta, c\delta, p_1, p_2)$ -sensitive if for any $x, y \in \mathbb{R}^d$ it satisfies the following properties:*

- (i) If $\|x - y\|_2 \leq \delta$ then $\Pr_{h \in H}[h(x) = h(y)] \geq p_1$.
- (ii) If $\|x - y\|_2 \geq c\delta$ then $\Pr_{h \in H}[h(x) = h(y)] \leq p_2$.

LEMMA 3.1. (ANDONI & INDYK [AI06]) *For any “scale” $\delta > 0$, dimension $d > 0$, and $c > 1$, there exists a $(\delta, O(c)\delta, 1/n^{1/c^2}, 1/n^3)$ -sensitive family of hash functions for \mathbb{R}^d .*

THEOREM 3.1. *Let X be a set of n points in Euclidean space, with spread Δ , and let $c > 1$. Then, there exists a 2-hop spanner for $(X, \|\cdot\|_2)$ with dilation $O(c)$, and with $O(n^{1+1/c^2} \log n \log \Delta)$ edges.*

Proof. We perform the same construction as in Theorem 2.1, but instead of Lipschitz partitions we use the family of hash functions given by Lemma 3.1. Specifically, a random hash function h chosen from H induces a partitioning that, by the property (ii) of the LSH functions, guarantees that the diameter of each cluster in the partition is at most $O(c\delta)$ with probability $1 - 1/n$. By selecting $k = n^{1/c^2} \log n$ partitions for each of the $O(\log \Delta)$ scales, and proceeding exactly as in the proof of Theorem 2.2, the theorem follows.

THEOREM 3.2. *Let X be a set of n points in Euclidean space, and let $c > 1$. Then, there exists a spanner for $(X, \|\cdot\|_2)$ with dilation $O(c)$, and with $O(n^{1+1/c^2} \log n \log c)$ edges.*

Proof. Similarly to the proof of Theorem 3.1, we perform the same construction as in Theorem 2.2, but instead of Lipschitz partitions we use the family of hash functions given by Lemma 3.1. We select $k = n^{1/c^2} \log n$

partitions for each of the $O(\log \Delta)$ scales, and we proceed as in the proof of Theorem 2.2. The only modification needed is that before performing LSH at each scale δ , we place into singleton clusters all points x , such that the ball of radius $O(c\delta)$ around x contains only x . We remove all these points that end up in singleton clusters, and we perform LSH on the remaining subset. Taking the union of the LSH partition, together with all singleton clusters, gives the desired partition. This modification guarantees that in the auxiliary tree T (as in the proof of Theorem 2.2), for every branch B of length L , with all vertices having a single child, in all but the $O(\log c)$ top levels, the point x that corresponds to B appears in a singleton cluster; therefore, it does not contribute any edges to the spanner at these levels. The rest of the analysis is exactly the same.

4 Lower bound for 2-hop spanners

We will show a dilation/size trade-off for 2-hop Euclidean spanners (i.e. such that for every pair of points x, y , there exists a low-dilation path in the spanner between x and y , that contains at most two edges).

We will use the point set $X = \{0, 1\}^d$, so $n = |X| = 2^d$.

Let $0 < \alpha < \beta < 1/2$, be parameters that can be optimized later. Let $r = \alpha d$, and $R = \beta d$. Assume w.l.o.g. that r is an odd integer.

Let G be a 2-hop spanner for the metric $(\{0, 1\}^d, \ell_2)$. Suppose that G has dilation at most $\sqrt{\beta/\alpha}$. As usual, we may assume that every edge $\{x, y\} \in E(G)$ has length $\|x - y\|_2$.

For every $x \in X$, let

$$S_x = \{x\} \cup \{y \in X : \|x - y\|_1 \leq R, \text{ and } \{x, y\} \in E(G)\}.$$

LEMMA 4.1. *For any $x \in X$, we have $|S_x| \leq ne^{-d(\frac{1}{2}-\beta)^2}$.*

Proof. Following Motwani *et al.* [MNP07] we have

$$|S_x| \leq \sum_{i=0}^R \binom{d}{i} = \sum_{i=0}^{\beta d} \binom{d}{i} \leq 2^d e^{-d(\frac{1}{2}-\beta)^2}.$$

For any $x \in X$, let

$$N_x = \{y \in X : \|x - y\| \leq r\}.$$

LEMMA 4.2. *For any $x \in X$, we have $N_x \subseteq \bigcup_{y \in X: x \in S_y} S_y$.*

Proof. Since G is a 2-hop spanner of dilation at most $\sqrt{\beta/\alpha}$, it follows that for every $z \in N_x$, either $z \in S_x$, or there exists $y \in X \setminus \{x\}$, with $\{x, y\} \in E(G)$, and $\{z, y\} \in E(G)$, and such that $\|x - y\|_1 \leq R$, and $\|z - y\|_1 \leq R$. Thus, for every $z \in N_x$, there exists $y \in X$, with $\{x, z\} \subseteq S_y$, which implies the assertion.

LEMMA 4.3. ([MNP07]) *Let r be an odd integer, and let $B \subseteq \{0,1\}^d$, with $B \neq \emptyset$. Consider the random variable Q_B defined as follows: Choose $z \in B$ uniformly at random, and perform r steps of the standard random walk on the Hamming cube starting from z . Let Q_B be the final point. Then,*

$$\Pr[Q_B \in B] \leq \left(\frac{|B|}{2^d}\right)^{\frac{e^{2r/d}-1}{e^{2r/d}+1}}.$$

Following [MNP07], for a point $x \in X$, we define $W_r(x)$ to be the following random variable: Start from x , and perform r steps of the standard random walk on the Hamming cube. Let $W_r(x)$ be the final vertex.

For a pair of points $x, y \in X$, define

$$p_{x,y} = \begin{cases} \Pr[W_r(x) \in S_y] & \text{if } x \in S_y \\ 0 & \text{if } x \notin S_y \end{cases}$$

LEMMA 4.4. *For any $x \in X$, we have $\sum_{y \in X} p_{x,y} \geq 1$.*

Proof. We have

$$\begin{aligned} \sum_{y \in X} p_{x,y} &= \sum_{y \in X: x \in S_y} \Pr[W_r(x) \in S_y] \\ &\geq \Pr \left[W_r(x) \in \bigcup_{y \in X: x \in S_y} S_y \right] \\ (4.4) \quad &= \Pr \left[W_r(x) \in N_x \cap \left(\bigcup_{y \in X: x \in S_y} S_y \right) \right] \\ (4.5) \quad &= \Pr[W_r(x) \in N_x] \\ (4.6) \quad &= 1, \end{aligned}$$

where (4.4) & (4.6) follow as $\Pr[W_r(x) \in N_x] = 1$, and (4.5) by Lemma 4.2

THEOREM 4.1. *Let G be a 2-hop spanner for $(\{0,1\}^d, \ell_2)$, with no Steiner nodes. Let $0 < \alpha < \beta < 1/2$, such that αd is an odd integer. Suppose that the dilation of G is at most $\sqrt{\beta/\alpha}$. Then,*

$$|E(G)| \geq \frac{1}{2} n e^{d(\frac{1}{2}-\beta)^2 \frac{e^{2\alpha}-1}{e^{2\alpha}+1}}.$$

Proof. Let $\lambda = \frac{e^{2\alpha}-1}{e^{2\alpha}+1}$, and let $M = n e^{-d(\frac{1}{2}-\beta)^2}$. We have that for any $y \in X$,

$$\begin{aligned} \mathbb{E}_{x \in S_y} [p_{x,y}] &= \Pr_{x \in S_y} [W_r(x) \in S_y] \\ &= \Pr[Q_{S_y} \in S_y] \\ (4.7) \quad &\leq \left(\frac{|S_y|}{n}\right)^\lambda \end{aligned}$$

$$(4.8) \quad \leq \left(\frac{M}{n}\right)^\lambda.$$

where (4.7) follows by Lemma 4.3, and (4.8) by Lemma 4.1. We have

$$\begin{aligned} (4.9) \quad n &\leq \sum_{y \in X} \sum_{x \in X} p_{x,y} \\ &= \sum_{y \in X} |S_y| \cdot \mathbb{E}_{x \in S_y} [p_{x,y}] \\ (4.10) \quad &\leq \sum_{y \in X} |S_y| \frac{M^\lambda}{n^\lambda}, \end{aligned}$$

where (4.9) follows by Lemma 4.4, and (4.10) by (4.8). By (4.10) we have

$$\begin{aligned} \sum_{y \in X} |S_y| &\geq \frac{n^{1+\lambda}}{M^\lambda} \\ &= n e^{d(\frac{1}{2}-\beta)^2 \lambda} \\ &= n e^{d(\frac{1}{2}-\beta)^2 \frac{e^{2\alpha}-1}{e^{2\alpha}+1}}. \end{aligned}$$

Since $|E(G)| \geq \frac{1}{2} \sum_{y \in X} |S_y|$, the result follows.

COROLLARY 4.1. *For any $c \geq 1$, there exists $\gamma = \gamma(c) = \Omega(1/c^2)$ satisfying the following. Let $d \geq 1$, and let G be a 2-hop spanner for $(\{0,1\}^d, \ell_2)$, with no Steiner nodes, and with dilation at most c . Then, $|E(G)| = \Omega(n^{1+\gamma})$.*

Proof. Set $\beta = 1/4$, $\alpha = 1/(4c^2)$. Theorem 4.1 yields the dilation $\sqrt{\beta/\alpha} = c$, and the lower bound of

$$\begin{aligned} |E(G)| &\geq n e^{d(\frac{1}{2}-\beta)^2 \frac{e^{2\alpha}-1}{e^{2\alpha}+1}} \\ &> n^{1+\log_2 e(1/4-1/2)^2 \frac{1+2\alpha-1}{e^{2\alpha}+1}} \\ &= n^{1+\Omega(1/c^2)}. \end{aligned}$$

References

- [AI06] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Proc. 47th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 459–468, 2006.
- [CCG⁺98] M. Charikar, C. Chekuri, A. Goel, S. Guha, and S. A. Plotkin. Approximating a finite metric by a small number of tree metrics. In *Proc. 39th Annu. IEEE Sympos. Found. Comput. Sci.*, pages 379–388, 1998.
- [Epp00] D. Eppstein. Spanning trees and spanners. In J.-R. Sack and Jorge Urrutia, editors, *Handbook of Computational Geometry*, pages 425–461. Elsevier, 2000.
- [Erd63] P. Erdős. Extremal problems in graph theory. In *Theory Graphs Appl., Proc. Symp. Smolenice*, pages 29–36, 1963.

- [HIM12] S. Har-Peled, P. Indyk, and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. *Theory Comput.*, page to appear, 2012.
- [HM06] S. Har-Peled and M. Mendel. Fast construction of nets in low dimensional metrics, and their applications. *SIAM J. Comput.*, 35(5):1148–1184, 2006.
- [JL84] W. B. Johnson and J. Lindenstrauss. Extensions of lipschitz mapping into hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.
- [MNP07] R. Motwani, A. Naor, and R. Panigrahy. Lower bounds on locality sensitive hashing. *SIAM J. Discrete Math.*, 21(4):930–935, 2007.
- [NS07] G. Narasimhan and M. Smid. *Geometric spanner networks*. Cambridge University Press, 2007.
- [PS89] D. Peleg and A. Schäffer. Graph spanners. *J. Graph Theory*, 13:99–116, 1989.
- [Sal91] J. S. Salowe. Construction of multidimensional spanner graphs, with applications to minimum spanning trees. In *Proc. 7th Annu. ACM Sympos. Comput. Geom.*, pages 256–261, 1991.
- [Vai91] P. M. Vaidya. A sparse graph almost as good as the complete graph on points in k dimensions. *Discrete Comput. Geom.*, 6:369–381, 1991.
- [Zwi01] U. Zwick. Exact and approximate distances in graphs - a survey. In *Proc. 9th Annu. European Sympos. Algorithms*, pages 33–48, 2001.