On Nearest-Neighbor Graphs

David Eppstein [1] *Michael S. Paterson* [2] *Frances F. Yao* [3]

January 12, 2000

Abstract

The "nearest neighbor" relation, or more generally the "k nearest neighbors" relation, defined for a set of points in a metric space, has found many uses in computational geometry and clustering analysis, yet surprisingly little is known about some of its basic properties. In this paper, we consider some natural questions that are motivated by geometric embedding problems. We derive bounds on the relationship between size and depth for the components of a nearest-neighbor graph and prove some probabilistic properties of the k-nearest-neighbors graph for a random set of points.

1 Introduction

Neighborhood-preserving mappings from a set of points in space to a regular data array are useful for speeding up many computations in physical simulation. For example, complicated many-body interactions between particles can often be approximated by the dominant forces exerted by near neighbors. If the data for nearby particles is stored with close address indices in an array, one can take advantage of fast vectorized operations in contiguous memory to perform such computations. The question is, therefore, whether such neighborhood-preserving mappings from a point set to a regular array always exist. In this paper, we attempt to answer this question by first studying the performance of a known mapping scheme, called the monotone logical grid (MLG), and then establishing some combinatorial properties of the nearest-neighbor graph that are relevant to geometric embedding.

The outline of the paper is as follows. In Section 2, we introduce some basic concepts and notation. In Section 3, we describe the MLG mapping and analyze its performance. We give in Section 4 some examples of point sets which are hard to embed in a grid. In Sections 5, 6, and 7 we prove a polynomial upper bound on the size of a component of the NNG in terms of its diameter. We discuss higher dimensional generalizations of our bounds in Section 8. In Section 9 we give some concluding remarks and open problems.

A preliminary version of this paper was presented at ICALP'92 [14].

[1]Dept. of Information and Computer Science, University of California, Irvine, CA 92697, USA, http://www.ics.uci.edu/~eppstein/, eppstein@ics.uci.edu. Supported in part by NSF grant CCR-9258355 and matching funds from Xerox Corp.

[2]Dept. of Computer Science, University of Warwick, Coventry, CV4 7AL, England, Mike.Paterson@dcs.warwick.ac.uk. The work was done while this author was visiting Xerox Palo Alto Research Center. He is partially supported by the ESPRIT BRA Program of the EC under contract 7141 (ALCOM II) and a grant from the Xerox Corporation.

[3]Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, USA, yao@parc.xerox.com.

2 Preliminaries

Let $V = \{v_1, v_2, \ldots, v_n\}$ be a set of points in R^ℓ. The *nearest neighbor* of v_i is a point v_j, $j \neq i$, with minimum Euclidean distance from v_i. To make the nearest neighbor unique we choose the point v_j with maximum index in case of ties, and denote it by $\mathrm{nn}(v_i)$. For any v, we define the directed edge $e(v) = \langle v, \mathrm{nn}(v)\rangle$. The *nearest-neighbor graph* of V, denoted by $\mathrm{NNG}(V)$, is the directed graph $\langle V, E\rangle$ where $E = \{e(v) | v \in V\}$. It is easy to verify that the graph $\mathrm{NNG}(V)$ has the following properties:

1. Along any directed path in $\mathrm{NNG}(V)$, the edges have non-increasing lengths.
2. The only cycles in $\mathrm{NNG}(V)$ are 2-cycles. For $|V| \geq 2$, each weakly connected component C of $\mathrm{NNG}(V)$ contains exactly one 2-cycle. This pair of vertices is called the *bi-root* of C.
3. For a point set V in two dimensions, $\mathrm{NNG}(V)$ is a planar graph. Furthermore, two edges incident at a vertex v must meet at an angle of at least 60°, hence the degree of a vertex is at most six. For point sets in general position, this degree bound can be reduced to five.
4. $\mathrm{NNG}(V)$, when considered as an undirected graph with the biroot treated as a single edge, is a subgraph of $\mathrm{DT}(V)$ (the Delaunay triangulation of V) and of $\mathrm{MST}(V)$ (the minimum spanning tree of V).

The degree bound in (3) also holds for minimum spanning trees. Monma and Suri [13] showed that, conversely, any tree with vertex degree at most five is the minimum spanning tree of some point set; thus minimum spanning tree topologies of general position point sets are exactly characterized by their degrees. (See [10] for complications arising from special position.) We show that a similar degree-based characterization does not work for nearest neighbor graphs: if $\mathrm{NNG}(V)$ has a tree with many vertices, it must contain a long path.

By (4), the nearest-neighbor graph in two dimensions can be constructed in asymptotically the same time as $\mathrm{DT}(V)$, i.e., $O(n \log n)$ for n points (see [11]). For general dimension ℓ, $O(n \log n)$ is also possible, but with a constant depending exponentially on the dimension [6, 16].

We can generalize $\mathrm{NNG}(V)$ to k-$\mathrm{NNG}(V)$, the *k-nearest-neighbors* graph of V, by introducing k edges from a vertex to its k nearest neighbors. In any constant dimension ℓ, one can compute k-$\mathrm{NNG}(V)$ in time $O(kn \log n)$ [16] or even $O(kn+n \log n)$ [4, 5, 9]. The k-nearest-neighbors graphs are useful for certain clustering problems [12]. However at present they have not been studied extensively, and few of their combinatorial properties are known.

3 Monotone Logical Grid

Boris [3] proposed a data structure, called the *Monotone Logical Grid* (MLG), as a way of storing and indexing a set of points in R^ℓ for n-body simulation. (For an alternate approach to n-body simulation based on hierarchical clustering, see [4, 5].) The MLG maps a set S of n points in R^ℓ to an ℓ-dimensional array A of size $n^{\frac{1}{\ell}} \times \cdots \times n^{\frac{1}{\ell}}$; we assume for simplicity that $n^{\frac{1}{\ell}}$ is an integer. When

$\ell = 1$, the MLG is simply a sorted linear array. For $\ell \geq 2$, we first sort the points into $n^{\frac{1}{\ell}}$ buckets of equal size by the values of their x_1-coordinates. Then the i^{th} bucket is stored recursively as an $(\ell - 1)$-dimensional MLG in the subarray of A corresponding to $x_1 = i$.

It was observed in [3] that, on average, the MLG seems to preserve near-neighbor relations rather well. In other words, points that are close in real space tend to have close addresses in the MLG array. However, these claims were based on experimental data and no precise statements or proofs were given. By analyzing the performance of MLG for a random point set, we can indeed make the following statement in support of the empirical results.

Theorem 1. *Let V be a set of m^2 points chosen independently and uniformly within a square of size $m \times m$. Let k be a positive integer and ϵ a positive real number. For any $v \in V$, with probability at least $1 - \epsilon$, all k nearest neighbors of v lie within x-offset $O(\sqrt{k})$ and y-offset $O(\sqrt{m \log k})$ from the address of v in the MLG for V.*

Proof: Let (a, b) be the coordinates of v with respect to the $m \times m$ square M in R^2, and let (A, B) be the address assigned to v in the MLG, where $1 \leq A, B \leq m$.

We will choose z such that with high probability all k nearest neighbors of v lie within the region Q defined by the intersection of M with the $2z \times 2z$ square centered at (a, b). (See Figure 1.) Let S_j be the set of points assigned to the j^{th} column of the MLG. We choose also an integer w sufficiently large that with high probability the set of points in Q is contained in the set $\bigcup_{A-w \leq j \leq A+w} S_j$, i.e., the $2w + 1$ columns of the MLG centered about v.

Let T_j be the subset of S_j consisting of those points lower than Q, i.e., with y-coordinate in M less than $b - z$. We will show that, with high probability, there are at least $b - z - \Theta(\sqrt{m \log k})$ points in T_j and so any point in Q in the j^{th} column of the MLG has y-address at least this large.

The following Chernoff-type result is used to justify each of our probability claims. (See, e.g., [1].)

Proposition 1. *Let X be a sum of n independent random variables and let X_0 be the expectation of X. Then for any $a > 0$:*

$$\Pr[X < X_0 - a] < e^{\frac{-a^2}{2X_0}}, \tag{1}$$

$$\Pr[X > X_0 + a] < e^{\frac{-a^2}{2X_0} + \frac{a^3}{2X_0^2}}, \tag{2}$$

$$\Pr[X > X_0 + a] < e^{\frac{-2a^2}{n}}. \tag{3}$$

Choose $\delta > 0$. (The ϵ in the statement of the Theorem will be a small multiple of δ.) The part of M within distance z of (a, b) has area at least $\frac{1}{4}\pi z^2$ and is contained in Q. By Proposition 1(1) the probability that this part of M contains at least k points other than v is more than $1 - \delta$ for some choice of $z = \Theta(\sqrt{k})$. Hence, with at least this probability, Q contains the k nearest neighbors of v.

The vertical strip of M defined by $a \leq x \leq a + z$ has area at most zm and so, with probability at least $1 - \delta$, contains at most wm points where, by Proposition 1(2), we may choose $w = z + \Theta(\sqrt{z}) = \Theta(\sqrt{k})$. If this strip contains at most wm points then these are contained in at most $w + 1$ columns of the MLG. The same argument holds for a similar strip to the left of v. Thus, with probability at least $1 - 3\delta$, every k-neighbor of v has x-address differing by at most $w = \Theta(\sqrt{k})$ from A.

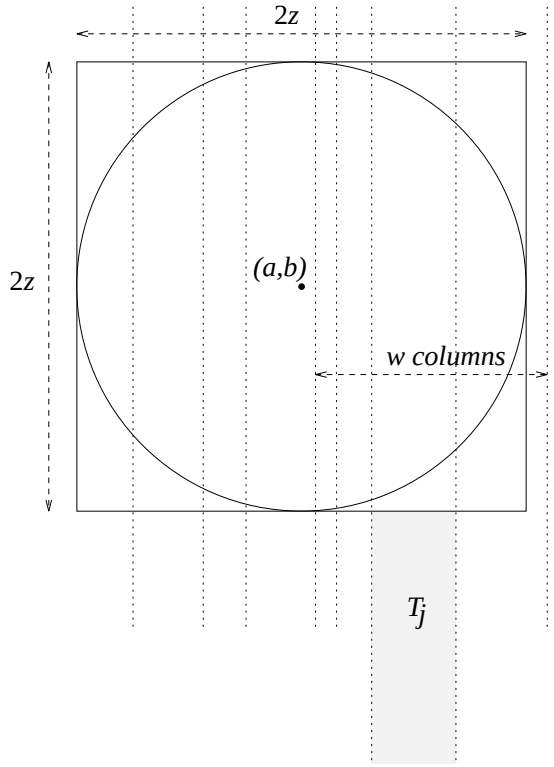

Figure 1. Mapping the neighborhood of v to the MLG.

To prove the bound on the y-offsets of k-neighbors we make use of the fact that the y-coordinates of the m points corresponding to any column of the MLG are uniformly and independently distributed over the real interval $[0, m]$.

We want to find some suitable d such that the probability is at least $1 - \delta/(2w+1)$ that $|T_j| > b - z - d$. We may assume that $b \geq z$ since otherwise the result is trivial. The height of the part-strip defining the set T_j is $b-z$. By simple estimation from Proposition 1(1), we can take $d = \Theta(\sqrt{m \log k})$.

This result holds for all $A - w \leq j \leq A + w$ and so, with probability at least $1 - \delta$, all points in Q have y-address greater than $b - z - d$. A similar bound for the region above Q shows that with the same probability all points in Q have y-address less than $b + z + d$. To complete the estimation of the maximal y-offset we need only show that $|B - b|$ is probably small. Indeed, Proposition 1(1) and (3) confirm that, with probability at least $1 - \delta$, $|B - b| < d'$ for some $d' = \Theta(\sqrt{m})$.

Combining all these probabilities, we find that, with probability at least $1 - 6\delta$, our choices of q,w,d and d' all have the desired properties and guarantee that the y-offsets of all k-neighbors of v are at most $\Theta(\sqrt{m \log k})$. □

In spite of this good expected performance, it is not hard to construct examples where the pairs of endpoints of almost all edges in $\mathrm{NNG}(V)$ are placed far apart in the MLG.

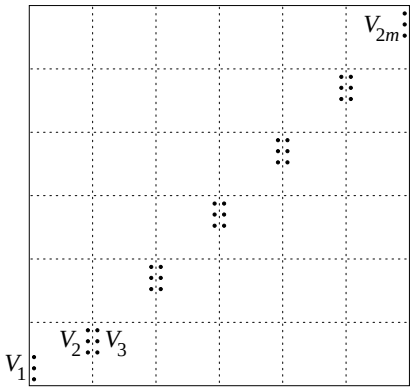

Figure 2. Monotone Logical Grid with large dilation.

Example 1. *(See Figure 2.) Let V consist of $2m$ subsets $V_1, V_2, \ldots, V_{2m}$, where each subset has $m/2$ points, and let ϵ be chosen such that $1/(2m) > \epsilon > 0$. For $i \geq 1$, the points of V_{2i} are equally spaced on the vertical segment $(i-\epsilon) \times [i-\frac{1}{2}, i]$, and the points of V_{2i+1} are equally spaced on the adjacent vertical segment $(i+\epsilon) \times [i-\frac{1}{2}, i]$. For V_1, we place its points uniformly on the interval $\epsilon \times [0, \frac{1}{2}]$). Thus, for $1 \leq i \leq m-1$, the points of V_{2i} and V_{2i+1} are matched pairwise as nearest neighbors. However, under the MLG mapping, the points of V_{2i} will have y-addresses $m/2+1, m/2+2, \ldots, m$, while the points of V_{2i+1} will have y-addresses $1, 2, \ldots, m/2$. Thus the y-offset in the MLG for nearest neighbors is $m/2$ for all points except those in V_1 and V_{2m}.*

One natural strategy for embedding NNG(V) into a grid is to embed the individual components of NNG(V) separately. This could certainly work well if all components of NNG(V) were small. Some experimental results [15] on pseudo-random sets of 10,000 and 25,000 points yield the distributions of component size shown in Table 1. We may therefore expect that a random set will have very few large components, though we have, at present, little exact knowledge regarding the distribution of component size for a random NNG(V). We can however prove the following result on the expected number

	Number of components of indicated size										
$\|V\|$	2	3	4	5	6	7	8	9	10	11	*Total number*
10000	1159	873	563	250	145	60	21	9	1	1	308
25000	2887	2336	1360	688	310	126	57	11	3	1	7779

Table 1. Distributions of component size in NNG's

of components in NNG(V) (cf [2], Lemma 1). Note that this result agrees closely with the experimental results shown in the Table. Our proof involves counting the number of repeated edges in NNG(V); several similar problems of counting edges in geometric graphs have been studied by Devroye [8].

Theorem 2. *The expected number of components in NNG(V) for a uniform random point set V in the unit square is asymptotic to approximately* $0.31 \times |V|$.

Proof: The expected number of components is half the expected number of elements which are in a bi-root of some component. In the following lemma we calculate the latter quantity for a Poisson distribution in the plane. If we consider the restriction of such a distribution to the unit square, the expected number of elements in the square whose nearest neighbor is altered by the restriction is only $O(\sqrt{|V|})$. The Theorem therefore follows directly. □

Lemma 1. *For a Poisson distribution of points in the plane, the probability that a given point is in a 2-cycle of the NNG, i.e., is in the bi-root of its component, is*

$$\frac{6\pi}{8\pi + 3\sqrt{3}} \approx 0.6215.$$

Proof: For such a distribution, the probability that any given region with area A is empty is k^A for some fixed k. A point p is in a 2-cycle if and only if the nearest neighbor q of p has p as *its* nearest neighbor, i.e., $C_p \cup C_q$ contains only the points p and q, where C_p and C_q are the circles with radius $r = ||p - q||$ centered at p and q respectively. Let $E(x) = k^{\pi x^2}$ be the probability that a given circle of radius x is empty. Since the random variable r is the maximum value such that the interior of C_p is empty (except for p), the density distribution of r is $-\frac{d}{dr}E(r)$. The probability that q has p as its nearest neighbor is the probability that $C_q \setminus C_p$ is empty. By simple geometry, the area of this region is $c\pi r^2$, where $c = 1/3 + \sqrt{3}/(2\pi)$, so the probability is $k^{c\pi r^2} = E(r)^c$ as a function of r. Hence the required probability is

$$-\int_{r=0}^{\infty} E(r)^c \frac{d}{dr}E(r)dr = -\int_{r=0}^{\infty} E(r)^c dE(r) = \frac{-1}{1+c}\left[E(r)^{1+c}\right]_{r=0}^{\infty} = \frac{1}{1+c} = \frac{6\pi}{8\pi + 3\sqrt{3}}.$$

□

This proof generalizes immediately to any fixed dimension d. The constant obtained depends only on the corresponding ratio c_k arising from the intersection of two unit balls. We have $c_1 = 1/2$, $c_2 \approx 0.61$, $c_3 = 0.6875$, and $c_4 \approx 0.75$, and $c_k \to 1$ as $k \to \infty$.

4 Dilation of Embeddings for the NNG

The *dilation* of an embedding of a k-NNG in a graph G is the maximum over all edges $\langle u, v\rangle$ in the k-NNG of the path length in G between the images of u and v. It is easy to see that any embedding of a rapidly branching tree in a finite-dimensional array must have a large dilation. We begin by constructing a 2-NNG which contains such a tree.

Example 2. *Consider Figure 3, where a recursively constructed tree layout is shown. Define the tree T_0 to be a single vertex. For $r > 0$, the tree T_r is constructed as follows. The root of T_r has two children, w_1 and w_2, placed with y-offset -2^r and x-offsets 2^r and -2^r respectively from the root. Each w_i is the root of a copy of T_{r-1}. It is clear from the diagram that the edges from any vertex v of T_r go to the nearest q neighbors of v for some $q \leq 2$. Thus T_r is a subgraph of the 2-NNG for the underlying point set.*

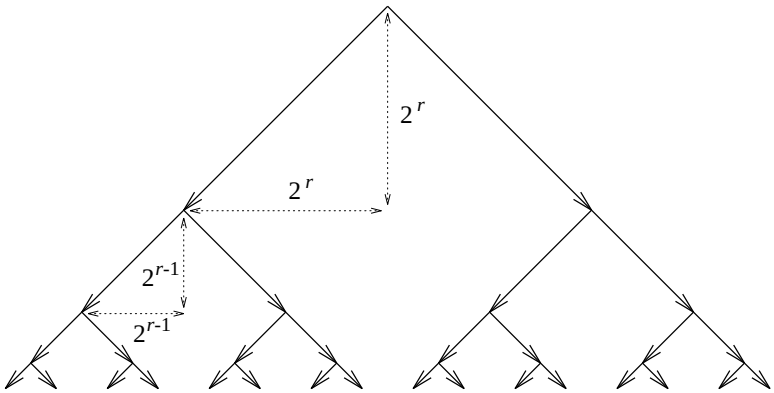

Figure 3. Embedding of the tree T_r ($r = 4$) in a 2-NNG.

Theorem 3. *For all n, there exists a planar set V of n vertices such that any embedding of 2-NNG(V) in an ℓ-dimensional array has dilation $\Omega(n^{1/\ell}/\log n)$. In particular, for $\ell = 2$ any embedding must have dilation $\tilde{\Omega}(n^{1/2})$.* [1]

Proof: For the 2-NNG in Example 2, the maximum (undirected) path length in the graph between any two vertices is $O(\log n)$. For any set of n elements of a ℓ-dimensional array, $\Omega(n^2)$ pairs of elements are Manhattan-distance $\Omega(n^{1/\ell})$ apart; some edge on the path connecting any such distant pair then gives the stated dilation. □

For an NNG, we define the *depth* $d(v)$ of a vertex v to be the path length in the NNG from v to a vertex in a bi-root. The *depth* of an NNG is the maximum depth of any vertex of the graph.

In order to show that a 1-NNG may require considerable dilation, we construct an example with depth $O(D)$ and $\Omega(D^5)$ vertices.

Example 3. *The graph is constructed using D^3 grids of D^2 vertices. Each grid is approximately square but the grids are of widely differing size. One such grid is shown within a dashed box in Figure 4. The construction proceeds in three further stages. In the first, we connect D of these grids to form a chain, as illustrated in Figure 4. The geometric sizes of the grids increase exponentially from left to right in this chain, with a ratio of D between successive grids. The edges within one grid are all approximately the same size, and each grid is connected into the chain by a path of D downward (270° to the x-axis) edges of about this same size. The next two edges to the left from the point of entry of this path are also of about this same size.*

After the first stage, if the exact layout is chosen with care, the result is an NNG of $\Theta(D^3)$ vertices, in the form of a chain of length $\Theta(D)$ with $\Theta(D)$ grids hanging from it. The depth is only $\Theta(D)$ at this point, but we now need the second stage.

The leftmost vertices of D chains as constructed in the first stage are connected using 90° edges to form an upward chain, as shown in Figure 5. The scales of successive members of the upward chain

[1] The tilde accent indicates the suppression of a polylogarithmic factor.

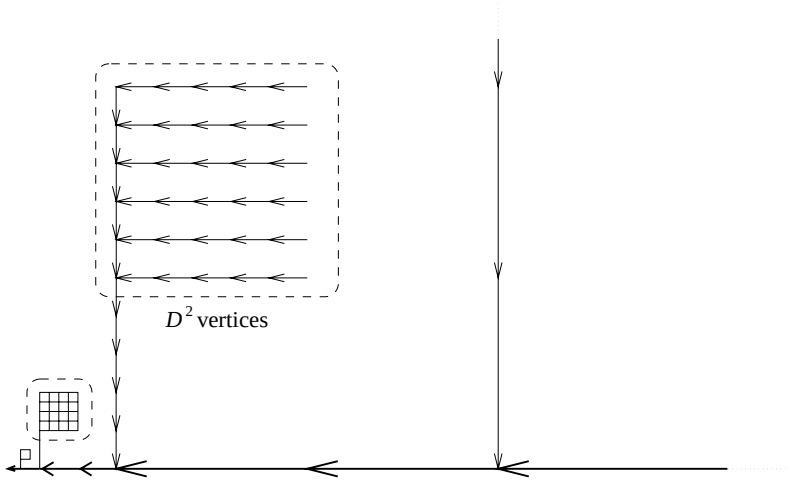

Figure 4. Chain of D D^2-vertex grids with connections via 180° edges.

differ extremely widely, each is about D^D times as large as its successor. This difference makes it impossible to give complete and accurate pictures of the construction. Our figures attempt to illustrate the ideas of the construction.

After this second stage, we have an NNG of $\Theta(D^4)$ vertices, consisting of a path of 90° edges, from which hang D subtrees, each of which is a chain of grids. The depth is still only $\Theta(D)$.

Each chain of the second stage ends with an upward edge to a final vertex. In the third stage, the final vertices of D of these chains are connected into a rightward chain using 0° edges, as shown in Figure 6. Note that each smaller second-stage chain fits into a "socket" provided at the end of the next larger chain. This completes the construction.

Lemma 2. *There is a connected 1-NNG with depth D and $\Omega(D^5)$ vertices.*

Proof: The number of vertices in the above construction is clearly $\Omega(D^5)$. The depth within each grid is $O(D)$. Each chain of each of the three types has length $O(D)$ and so the total depth is $O(D)$. By choosing constant factors in the construction appropriately we can make the depth exactly D.

To show that it is a 1-NNG one needs to verify that each vertex can be placed so that the edge depicted for it is to the actual nearest neighbor of the vertex. Note that, whenever there is a very long edge connecting to a smaller-scale structure, all the vertices in this structure are on the far side of a perpendicular line to the edge through its opposite endpoint. □

Theorem 4. *For any n, there exists a planar set V of n vertices such that any embedding of NNG(V) in an ℓ-dimensional array has dilation $\Omega(n^{\frac{1}{\ell}-\frac{1}{5}})$. In particular, for $\ell = 2$ we obtain a dilation of $\Omega(n^{3/10})$.*

Proof: The proof is similar to the proof of the previous theorem. In the NNG we constructed, the maximum (undirected) path length in the graph between any two vertices is $O(n^{1/5})$. The result follows

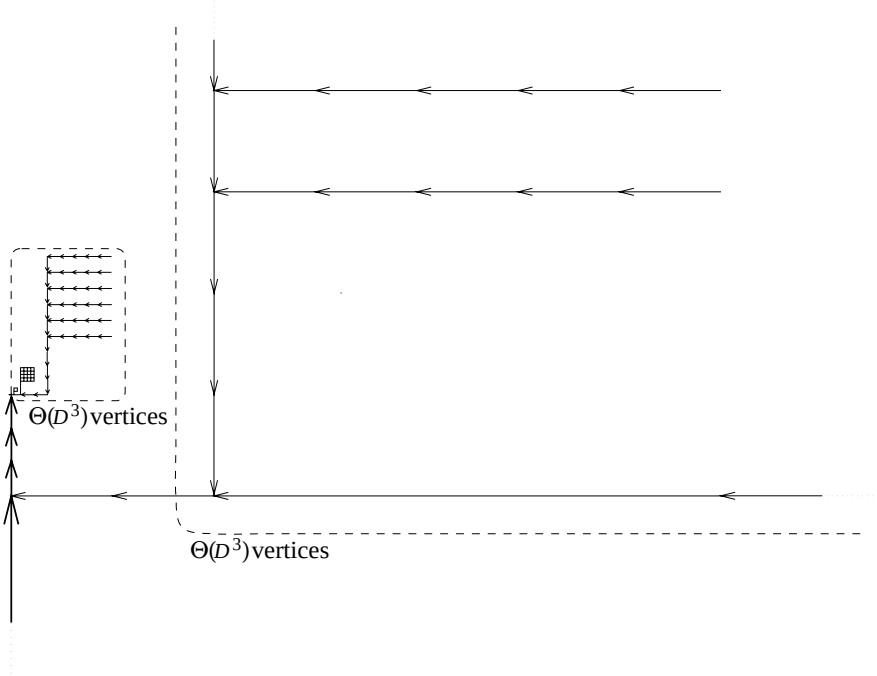

Figure 5. Chain of D $\Theta(D^3)$-vertex chains connected using 90° edges.

from comparing this length with the $\Omega(n^{1/\ell})$ distance achieved by almost all pairs of points in the array. □

5 Size and Depth of the NNG

In the following three sections we show that the $\Omega(D^5)$ size for an NNG of depth D proved above is tight: any connected NNG of depth D has $O(D^5)$ vertices. Such a result, while interesting on its own, could also be a first step towards an algorithm for embedding NNG's in arrays with low dilation.

As the proof is long and complicated, we split it into three parts which build on each other. We first outline a simplified version of the argument which suffices to prove an upper bound of $O(D^9)$. In the next section we improve this bound to $O(D^6)$, and finally we prove the $O(D^5)$ bound.

It is convenient to define the depth $d(e)$ of an edge $e = e(v)$ to be $d(v)$. The length of an edge $e = \langle u, v\rangle$ is denoted $||e||$ or $||u - v||$. For any $v \in V$, define $C(v)$ to be the open circle with center v and radius $||e(v)||/2$.

Lemma 3. *For all v_i, v_j, $C(v_i) \cap C(v_j) = \emptyset$.*

Proof: If $C(v_i) \cap C(v_j) \neq \emptyset$, then

$$||v_i - v_j|| < ||e(v_i)||/2 + ||e(v_j)||/2 \leq \max\{||e(v_i)||, ||e(v_j)||\}.$$

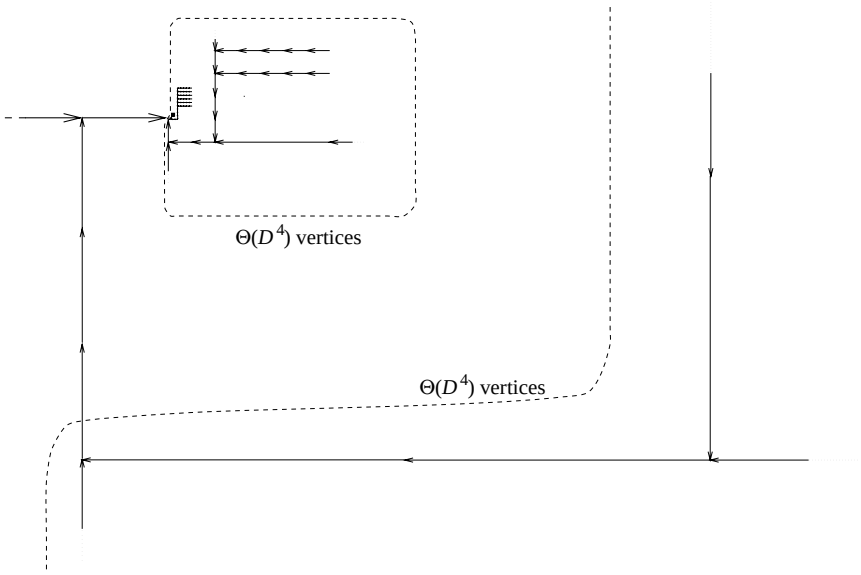

Figure 6. Chain of D $\Theta(D^4)$-vertex chains connected using 0° edges.

This inequality contradicts the nearest-neighbor property. □

Theorem 5. *If $G = \langle V, E\rangle$ is a connected plane NNG with depth D then $|V| = O(D^9)$.*

Proof: Let $R = \{z_1, z_2\}$ be the bi-root of G. We will normalize distances by taking $||z_1 - z_2|| = 1$. Let O be the midpoint of the edge (z_1, z_2). For $i \geq 0$, define B_i to be the open disk with radius 8^i and center O. For any vertex v, define $b(v)$ to be the minimal i such that B_i contains all vertices (including v) on the directed path from v to R, and let $S_i = \{v | b(v) = i\}$. The proof continues with three lemmas.

Lemma 4. $|S_i| = O(D^2)$.

Proof: Suppose $v \in S_i$. Since v or some vertex on the path from v to R is at a distance at least 8^{i-1} from the origin, and each edge on this path is of length at most $||e(v)||$, we have $||e(v)|| \geq (8^{i-1} - 1/2)/d(v) \geq (8^{i-1} - 1/2)/D$. The disk $C(v)$ is contained within a circle of radius $8^i + ||e(v)||/2$ centered at O, and $||e(v)|| \leq \min\{||v - z_1||, ||v - z_2||\} < 8^i + 8^{-i-1}$. By Lemma 3 the disks $C(v)$ are disjoint for all $v \in S_i$ and so a comparison of areas yields the inequality

$$((8^{i-1} - 1/2)/2D)^2 \cdot |S_i| \leq (8^i + (8^i + 8^{-i-1})/2)^2.$$

Hence, $|S_i| = O(D^2)$. □

Lemma 5. *For all $i > 0$, there are at most six directed edges from $\bigcup_{j>i} S_j$ to $\bigcup_{j<i} S_j$.*

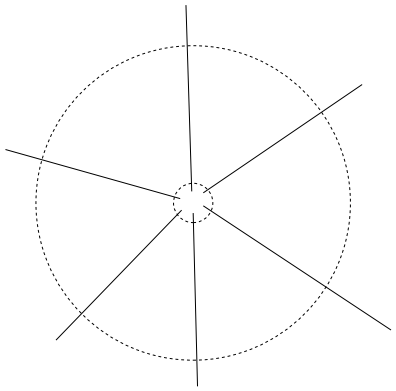

Figure 7. At most six edges cross from $\bigcup_{j>i} S_j$ to $\bigcup_{j<i} S_j$.

Proof: Any such edge comes from outside B_i to inside B_{i-1} (Figure 7). Consider any pair of edges $\langle v_1, w_1\rangle$, $\langle v_2, w_2\rangle$ such that vertices v_1, v_2 are outside B_i, while w_1, w_2 are within B_{i-1}, and $\angle v_1 O v_2 = \psi$ say. Without loss of generality we can assume that $||v_1 - O|| \geq ||v_2 - O||$. Suppose that $||v_1 - O|| = \theta t$ and $||v_2 - O|| = t$ where $\theta \geq 1$.

As w_1 is a nearest neighbor of v_1, we have

$$1 \leq \frac{||v_1 - v_2||}{||v_1 - w_1||} < \frac{t\sqrt{1 + \theta^2 - 2\theta\cos\psi}}{\theta t - 8^{i-1}} \leq \frac{\sqrt{1 + \theta^2 - 2\theta\cos\psi}}{\theta - 1/8},$$

since $t \geq 8^i$. Therefore

$$0 > \theta - 1/8 - \sqrt{1 + \theta^2 - 2\theta\cos\psi} \geq 7/8 - \sqrt{2(1 - \cos\psi)},$$

since the middle expression is monotonically increasing in θ. Hence $\cos\psi < 79/128$ and so $\psi > 2\pi/7$. The lemma follows since, if there were seven or more edges from $\bigcup_{j>i} S_j$ to $\bigcup_{j<i} S_j$, then there would be two outer vertices v_1, v_2 subtending an angle of $2\pi/7$ or less at the origin. □

The *extension* operation on a set E of edges consists of deleting one edge $e \in E$ and replacing it by the set of predecessor edges of e. We associate with any set E of edges the *characteristic* of E, denoted $\chi(E)$, which is the 7-tuple $\langle d_1, \ldots, d_7\rangle$ where d_i is the depth of the edge with the i^{th} smallest depth among edges of E. If E has fewer than seven edges then some of the last components of $\chi(E)$ take the value ∞. A simple property of extensions is stated in the following lemma.

Lemma 6. *If the set E' is the result of applying one or more extension operations to E then $\chi(E') \succeq \chi(E)$, where '$\succeq$' denotes lexicographic order.* □

Define $E_i = \{\langle v, v'\rangle | b(v) \geq i, b(v') < i\}$. Since E_{i+1} can be obtained by applying a sequence of extension operations to E_i, we have $\chi(E_{i+1}) \succeq \chi(E_i)$.

If $\chi(E_{i+1}) = \chi(E_i)$ then every edge represented in $\chi(E_{i+1})$ must also have been a member of E_i. By Lemma 5 there can be at most six edges in $E_i \cap E_{i+1}$, and so in this case the characteristic of E_{i+1}

has at most six finite components. Hence each edge of E_{i+1} is represented in $\chi(E_{i+1})$. The equality of the characteristics therefore implies that $E_i = E_{i+1}$. Hence there can be no edge from S_i to S_j for any $j < i$, and so $S_i = \emptyset$ in this case.

Thus at each step in the sequence $\chi(E_1), \chi(E_2), \ldots$, either the characteristic increases in the lexicographic order, or else remains the same in which case S_i is empty. Since the characteristic can take at most $O(D^7)$ values, by Lemma 4 the total number of vertices in G is $O(D^7 \cdot D^2)$, and the Theorem is proved. □

6 Improved NNG Depth Bounds

Our proof that there are at most $O(D^6)$ points in a nearest neighbor graph with depth D follows the same general outline as the simpler proof of the $O(D^9)$ bound above. In the previous section, we defined disks B_i as having radius 8^i. Here we let the radius of B_i be K^i for some larger parameter K to be determined later. We define $b(v)$ and S_i as before, and Lemma 4 holds as before. Let A_i refer to the annulus $B_i \setminus B_{i-1}$. We tighten Lemmas 5 and 6 to show that in fact only $O(D^4)$ of the sets S_i can be non-empty. Note that S_i can be non-empty only if annulus A_i contains at least one point v for which $\mathrm{nn}(v)$ is in B_{i-1}.

By the *angle* between two line segments we refer to one of the two angles formed at the crossing of the lines containing the segments. We will use this term only when each segment has its two endpoints in two different annuli A_i, so that the 'outer' and 'inner' endpoints of the segments are unambiguously defined. Of the two possible angles, we choose the one between the rays that contain the two outer endpoints. If the two lines containing the segments are parallel we say that the segments have angle either 0° or 180° as appropriate.

The following lemma tightens Lemma 5, which proved a $2\pi/7$ bound on the angle between segments crossing the annuli.

Lemma 7. *Let edges e and f of NNG(V) pass entirely across a given annulus A_i (so that one endpoint is outside B_i and the other is inside B_{i-1}). Then angle ef must be at least $60^\circ - O(1/K)$.*

Proof: Define the two parallel infinite wedges W_1 and W_2, with an opening angle of 120°, so that the outer endpoint of e is on the centerline of the two wedges, and the sides of W_1 are tangent to B_{i-1} while W_2 is separated from B_{i-1} by far enough so that the apex of W_2 is nearer than B_{i-1} to any point within W_2 (Figure 8).

Note that for both e and f, the outer endpoint must have the inner endpoint as nearest neighbor (rather than vice versa) since by assumption there are at least two points in ball $B_1 \subset B_{i-1}$ and any pair of points in B_{i-1} is closer to each other than to all points outside B_i for sufficiently large K.

If the outer endpoint of f is outside wedge W_1, then (since the inner endpoint is inside the wedge) e must form an angle of at least 60° with f. If the outer endpoint of f is inside wedge W_2, then the outer endpoints of e and f would be nearer to each other than the distance from the further of the two to the apex of W_2, and hence the edge between those endpoints would have been chosen as part of NNG(V), contradicting the assumption that e and f are in NNG(V).

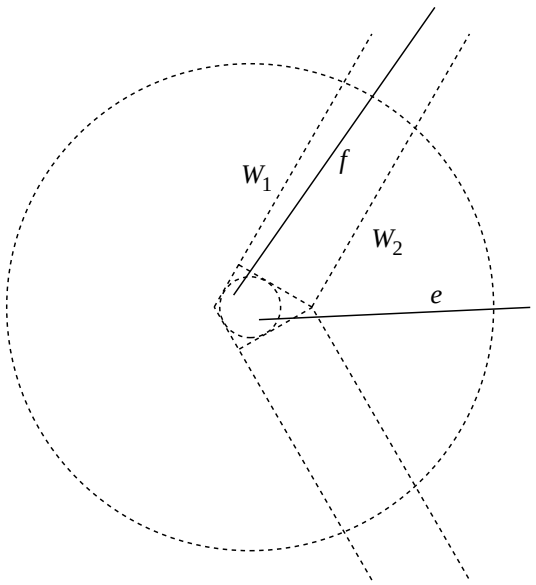

Figure 8. If e and f both cross A_i, then the smallest angle between them is formed when f is in a strip between two 120° wedges.

The remaining possibility is that f is in $W_1 \setminus W_2$. This region takes the form of two semi-infinite strips, with width proportional to the radius of B_{i-1}. Since f has length $\Omega(K)$ times that width, it must form an angle with e that is within $O(1/K)$ of 60°. □

Lemma 8. *Let e and f pass entirely across a given annulus A_i, and let the outer endpoints of e and f be in two non-adjacent annuli A_j and A_k, $k > j+1$. Then angle ef is at least $90^\circ - O(1/K)$.*

Proof: Let the outer endpoint of f be farther from O than the outer endpoint of e, and suppose the radius of B_{i-1} is r. Then $||f|| = r' = \Omega(K^3 r)$ and the circle of radius r' centered on the outer endpoint of f meets B_{i-1} but excludes the outer endpoint of e. Now the inner endpoint of e is in B_{i-1}, and $||e|| = \Omega(Kr)$ but $||e|| = O(r'/K)$. Hence the angle between e and f is at least $90^\circ - O(1/K)$. □

If we choose K large enough, the angles in the previous two lemmas will "look like" 60° and 90°, in that the following inequalities hold:

- $7(60^\circ - O(1/K)) > 360^\circ$, so that no annulus can be crossed by seven edges.
- $5(60^\circ - O(1/K)) + (90^\circ - O(1/K)) > 360^\circ$, so that if six edges cross an annulus, all angles are smaller than $90^\circ - \Omega(1/K)$.
- $3(90^\circ - O(1/K)) + 2(60^\circ - O(1/K)) > 360^\circ$, so that if there are five crossing edges, all but perhaps two angles are smaller than $90^\circ - \Omega(1/K)$.

Theorem 6. *If $G = \langle V, E \rangle$ is a connected plane NNG with depth D then $|V| = O(D^6)$.*

Proof: We consider two types of edge for each annulus A: those that cross the inner but not the outer boundary of A, and those that cross both boundaries. Two edges of the second type are *related in A* if they have an angle less than $90^\circ - \Omega(1/K)$ or if there is a sequence of related edges from one to the other. There can be at most three equivalence classes of related edges, except for the single case that four edges cross the annulus at approximate right angles.

As in the proof of our weaker bound, we label each annulus A by a characteristic $\chi(A)$. Here we modify $\chi(A)$ to consist of the 4-tuple of the four smallest depths of edges crossing the inner boundary of A, only allowing a single depth from each equivalence class of related edges. If there are not enough depths to fill out the tuple, we fill the remaining positions with the value ∞. Unlike the characteristics used in Lemma 6, these need not increase in lexicographic order. However we show that we can find a subsequence of the characteristic sequence, covering a constant fraction of the annuli A_i for which $S_i \neq \emptyset$, in which the characteristics do increase. Only $O(D^4)$ characteristics are possible in this subsequence, and therefore the entire sequence consists of $O(D^4)$ annuli.

Suppose we have constructed some such sequence out to some annulus A_i. Let X be the set of edges that cross both boundaries of A_i. We choose the next annulus to be the smallest annulus A_j, $j > i$, for which $S_j \neq \emptyset$, A_j does not contain an edge in X, and A_j is not crossed by four unrelated edges.

We now show that this selection process discards at most some constant number of the A_j for which $S_j \neq \emptyset$. Suppose we discard an annulus crossed by four unrelated edges, which must by Lemma 8 be at approximate right angles. No point can be within four circles having those edges as radii, which together cover all of the annulus except for a small region near the center. By an argument similar to that of Lemma 7, any edge crossing the outer annulus boundary would have angles greater than $60^\circ - O(1/K)$ with all four crossing edges. But no such angles, and hence no such edges, can exist. Thus the next larger annulus for which $S_j \neq \emptyset$ contains the endpoint of one of the four crossing edges. Such an annulus cannot itself be crossed by four unrelated edges, and so is not of this special form. So, as we search from A_i for the next annulus A_j, at least every alternate annulus we encounter is not of this form.

The remaining annuli discarded in the search from A_i for the next annulus A_j in the sequence each contain an endpoint of an edge in X, so there can be at most six such annuli. Thus after we try $O(1)$ annuli we will have run out of edges in X, and one of the next two annuli we encounter will be chosen as next in the sequence.

Finally, we show that the characteristic of the chosen A_j is larger than that of A_i. Let $\mathrm{E}(A_i)$ denote the set of edges corresponding to the characteristic $\chi(A_i)$. Let $X' \subseteq X$ be the set of edges that cross the inner boundaries of both A_i and A_j. As the next step in our proof, we show that there is some edge e in $\mathrm{E}(A_j) \setminus X'$. Since A_j contains no endpoint of an edge in X, each edge in X' crosses also the outer boundary of A_j. All its relatives in X are also in X', and are related in X'. But $|X' \cap \mathrm{E}(A_j)| \leq 3$ since we excluded from the sequence annuli with four unrelated crossing edges. If $\mathrm{E}(A_j)$ contains four unrelated edges, there is at least one such edge e not in X'. If $\mathrm{E}(A_j)$ contains fewer than four unrelated edges, every edge crossing the inner boundary of A_j must be in $\mathrm{E}(A_j)$, and in particular (since by assumption S_j is nonempty) there must be an edge $e = e(v)$, for v in S_j, which cannot be in X' since every edge in X' crosses the outer boundary of A_j.

In any case we have shown that some edge in $\mathrm{E}(A_j)$ is not also in X'. Let e be such an edge with the smallest depth. As we follow the path from e to the bi-root, we will eventually encounter an edge e' crossing the inner boundary of A_i. The depth of e' is strictly smaller than that of e. We form a characteristic χ' from an edge set E' constructed by adding e' to the set $X' \cap \mathrm{E}(A_j)$. We fill all remaining positions in the 4-tuple by ∞. Then since e' has lower depth than all edges in $\mathrm{E}(A_j) \setminus X'$, it follows that $\chi' < \chi(A_j)$ in the lexicographic order. On the other hand, all edges in $\mathrm{E}' \setminus \{e'\}$ are unrelated in A_i, and no relative of e' can be in X' or hence in E'. Thus χ' is a potential label for A_i and the true label $\chi(A_i)$ is no higher in lexicographic order.

We have shown that $\chi(A_i) < \chi(A_j)$. The subsequence we construct, and thus the entire sequence of annuli A_i for which $S_i \neq \emptyset$, has at most $O(D^4)$ members. Each set S_i contains $O(D^2)$ points by Lemma 4, and the theorem is proved. □

7 Tight NNG Depth Bounds

We now add further complications to the argument above, to remove a final factor of D and arrive at an $O(D^5)$ bound on the size of a depth-D NNG, matching the lower bound of Theorem 4.

Recall that we defined a sequence of $O(D^4)$ annuli A_i, each of which could contain $O(D^2)$ points. Both of these bounds are tight individually. However we can improve our bounds by proving that it is not possible for all annuli to contain many points.

We define a new characteristic $\xi(A)$ of an annulus A to be the 3-tuple of the three smallest depths of edges crossing the inner boundary of A, as in the definition of $\chi(A_i)$ allowing a single depth from each equivalence class of related edges. We define a sequence of *special annuli* as follows.

Suppose we have constructed such a sequence out to some annulus A_i. Let X be the set of edges chosen for $\xi(A_i)$. Let A_j be the next annulus for which some edge in X and all relatives of this edge do not cross A_j. We choose A_j as the next special annulus.

Lemma 9. *There are $O(D^3)$ special annuli.*

Proof: There are $O(D^3)$ possible values for $\xi(A)$. As we progress from smaller to larger annuli in the special sequence, $\xi(A)$ can only decrease or remain constant in two circumstances: (1) It can decrease if some edge e, one of the three edges of $\xi(A)$, with depth x less than that of some other edge in $\xi(A)$, has a relative that terminates, producing new unrelated edges of depth $x+1$. But then, by Lemma 8, edge e itself will terminate within $O(1)$ annuli, leading either to an overall increase in $\xi(A)$ or to case (2) below. (2) It can remain constant if some edge, one of the three edges of $\xi(A)$, with depth at least that of the other edges in $\xi(A)$, terminates, but some unrelated edge has the same depth. Then $\chi(A)$ must be of the form (a, b, c, c) but because of the monotonic sequence shown to exist in Theorem 6, this can only happen $O(1)$ times for each such value. □

Lemma 10. *For any annulus A, there are $O(D)$ annuli A_i with $S_i \neq \emptyset$ separating A from a special annulus.*

Proof: This follows from the fact that we can go through at most $O(D)$ different values of $\chi(A)$ before $\xi(A)$ also changes. □

We say that i is a *special index* if A_i is a special annulus. By Lemma 9 there are $O(D^3)$ special indices.

We make one further distinction among our annuli: we say that an annulus A is *bad* if some two of the three edges defining $\xi(A)$ terminate within D levels of each other, and that A is *good* if it is not bad. A slight extension of Lemma 8 shows that the three edges defining $\xi(A)$ for a good annulus form angles that are at least $90^\circ - c^D$ for some $c < 1$.

Lemma 11. *There are $O(D^3)$ bad annuli.*

Proof: Let A be a bad annulus. By Lemma 10, A is within $O(D)$ annuli of the next larger special annulus.

Assume first that the two edges in $\xi(A)$ with endpoints at similar levels include the shortest of the three edges in $\xi(A)$. Then within $O(D)$ annuli from the next special annulus larger than A, two of the three edges of $\xi(A)$ change. By a similar argument to that of Lemma 9, this can happen $O(D^2)$ times.

In the remaining case, the two edges with endpoints at similar levels are the longest two edges in $\xi(A)$. After $O(D)$ special annuli (and hence $O(D^2)$ total levels), these edges will both terminate, and hence all three edges in $\xi(A)$ will have terminated. But this can only happen $O(D)$ times. □

Lemma 12. *Let A_i be a good annulus, and let the closest special index to i be j. Then there is a constant $c < 1$ such that A_i contains $O(D + c^{|i-j|}D^2)$ points.*

Proof: Let $k = |i - j|$. Then there must be three edges of NNG(V) with inner vertices within an annulus at least k levels smaller than A_i, and outer vertices within an annulus at least k levels larger than A_i.

Since these edges are in NNG(V), the circles with these edges as radii cannot contain any points. However these circles cover all but an exponentially small fraction of A_i. Since their outer vertices are far from A_i, the circle boundaries are close to straight lines perpendicular to the edges through their centers: they differ from those straight lines in the area they cut from A_i by an exponentially small amount (c_1^k for some $c_1 < 1$). As noted above, the angles between these three lines are at least $90^\circ - c_2^D$ for some $c_2 < 1$. The inner vertices of these edges are (relative to the size of A_i) within distance c_3^k of each other for some $c_3 < 1$.

Therefore the uncovered area of A_i is c_1^k plus a trapezoid-shaped region with base at most c_3^k and angle at most $2c_2^D$. We simplify this remaining region of A_i in which points may be contained, to a rectangle with side lengths (measured relative to the radius of A_i) $O(1)$ and $O(c^{\min(k,D)})$ for c chosen as some appropriate function of c_1, c_2, and c_3.

As in Lemma 4, we can find a collection of disjoint circles, centered on each point in A_i, with radius $O(1/D)$ relative to the radius of A_i. If $c^k > 1/D$, each circle covers $1/D^2$ units of the rectangle's area, and there are $O(c^k D^2)$ circles. If $c^k \leq 1/D$, each circle covers a $1/D$ fraction of the rectangle's area, and there are $O(D)$ circles. □

Theorem 7. *If $G = \langle V, E\rangle$ is a connected plane NNG with depth D then $|V| = O(D^5)$.*

Proof: By Lemma 11 the number of points in bad annuli is $O(D^5)$. As in Theorem 6 there are $O(D^4)$ annuli, so the $O(D)$ terms of Lemma 12 add to $O(D^5)$. The $O(c^k D^2)$ terms can be charged to the nearest special annulus; they add in a geometric series to $O(D^2)$ per special annulus for $O(D^5)$ total. □

8 NNG in Higher Dimensions

We can generalize Theorems 4 and 6 to prove that in any fixed dimension ℓ, the size of a connected NNG is bounded above and below by a polynomial in its depth D.

We first consider generalizations of the lower bound construction of Theorem 4. Recall that we connected a sequence of points, each containing a square grid of $\Omega(D^2)$ points, using edges in three of the four directions parallel to coordinate axes. The fourth direction was reserved for edges connecting each level to its grid in such a way that the grid did not interfere with the other edges.

Theorem 8. *In any constant dimension ℓ, and for any D, there is a connected NNG with depth D and $\Omega(D^{2\ell+1})$ vertices.*

Proof: We again form a sequence of levels, each containing a grid of $\Omega(D^\ell)$ points in the positive orthant. The levels are connected by edges in $\ell + 1$ directions parallel to coordinate axes, namely the directions of the negative orthant together with a single positive-orthant direction. The placement of the grid to avoid interfering with other edges and the details of the connections between levels are essentially the same as those in Figures 4, 5, and 6. □

We next consider upper bounds on the size of a depth-D NNG. The bound of Theorem 5 generalizes to $O(D^{\ell+\tau(\ell)+1})$ where $\tau(\ell)$ is the number of disjoint unit hyperspheres which can touch a given unit hypersphere, i.e., the so-called *kissing number* in dimension ℓ. Conway and Sloane [7] give a fascinating history of the kissing number question. Currently, the exact value of $\tau(\ell)$ is known only for dimensions $\ell = 1, 2, 3, 8, 24$. For general ℓ, there are exponential upper bounds and lower bounds $\alpha^\ell \leq \tau(\ell) \leq \beta^\ell$ where $\alpha \approx 2^{0.207}$ and $\beta \approx 2^{0.401}$.

For the purpose of proving upper bounds on NNG component size, it is more convenient to use the following equivalent definition of the kissing number: $\tau(\ell)$ is the number of points that can be placed on the surface of sphere in ℓ dimensions so that the angular separation between any two points is at least 60°. The corresponding quantity for angular separation at least 90° is simply 2ℓ, and we can use this quantity in place of $\tau(\ell)$ to generalize Theorem 6 and prove a bound of $O(D^{3\ell})$ on the size of a connected NNG, giving a bound only singly exponential in ℓ in place of the doubly exponential generalization of Theorem 5. The example of Theorem 8 shows that we are at least within a constant factor of the right exponent.

Theorem 9. *In any constant dimension ℓ, if $G = \langle V, E\rangle$ is a connected NNG with depth D then $|V| = O(D^{3\ell})$.*

Proof: The discussions in Sections 5 and 6 generalize from dimension two to higher dimensions without any difficulty, once we change "circles" to "hyperspheres", "disks" to "balls", and the "characteristic" of an edge set from a 4-tuple to a 2ℓ-tuple. Lemma 8 can be used to show that at most 2ℓ subsets of crossing edges can be separated by angles of $90^\circ - \epsilon$, and that this number is further reduced to $2l - 1$ except in the single case of 2ℓ approximately orthogonal edges. The constant K used in Section 6 for the definition of B_i must be enlarged so that the $O(1/K)$ term in Lemma 8 is small enough for the definition of "approximately orthogonal" above. Lemma 4 generalizes to a bound of $O(D^\ell)$ points per higher-dimensional annulus. The proofs of the lemmas go through with little modification. □

9 Conclusion and Open Problems

We have explored some of the combinatorial properties of k-NNG's. Whereas for $k \geq 2$, a k-NNG can have components of size exponential in the depth, for $k = 1$ the size is limited by a polynomial. We have also constructed NNG's which require dilation $n^{\Omega(1)}$ for embeddings into an array of any finite dimension.

For uniform random distributions of points in the plane, Theorem 2 gives the expected number of connected components of the NNG. It would be of interest to estimate some further properties of the NNG, such as the distribution of component sizes and the expected size of the largest component. Similar questions arise for the k-NNG and, in particular, we would like to know the minimum value of k (as a function of n) for which the k-NNG can be expected to have a 'giant' component. An understanding of the probabilistic properties of the k-NNG might be exploited in designing embeddings with improved computational characteristics.

Our knowledge of the combinatorial properties of NNG's is still very limited. We know of no graph-theoretic characterization of NNG's and suspect that deciding, for a given graph, whether or not it is the NNG for some set of points may be a hard problem. So far we know even less about k-NNG's for $k > 1$.

In Theorem 4 we proved a bound of $\Omega(n^{\frac{1}{\ell}-\frac{1}{5}})$ on the dilation of a 1-NNG embedded in an ℓ-dimensional array. When $\ell \geq 5$ this bound becomes trivial, and in fact the example used to prove this bound can be embedded in a 5-dimensional array with constant dilation. It remains open whether such good embeddings are possible in general, or more generally whether Theorem 4 is tight even for $\ell = 2$. Perhaps the techniques used in our upper bounds can be used to provide good array embeddings of 1-NNG's.

The gap between the lower bound of $\Omega(D^{2\ell+1})$ in Theorem 8 and the upper bound of $O(D^{3\ell})$ in Theorem 9 is relatively small but still remains open. It seems likely that some of the ideas from the tight bound of Theorem 7 could be extended to the higher dimensional case.

An understanding of the relationship between the dimension of a space and the combinatorial or statistical properties of the k-NNG's would be of use in classification theory. For example, if a set of points were representable as a fairly well-behaved distribution over a manifold of unknown dimension, could we estimate the dimension of the manifold from statistical properties of the k-NNG?

Acknowledgements

The authors wish to thank Marshall Bern, Dan Greene and Shang-Hua Teng for helpful discussions.

References

[1] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley-Interscience. New York. 1992.

[2] M. Bern. Two probabilistic results on rectilinear Steiner trees. *Algorithmica* 3 (1988) 191–204.

[3] J. Boris. A vectorized 'near neighbors' algorithm of order N using a monotonic logical grid. *Journal of Computational Physics* 66 (1986) 1–20.

[4] P. B. Callahan. Optimal parallel all-nearest-neighbors using the well-separated pair decomposition. *Proc. 34th IEEE Symp. Foundations of Computer Science* (1993) 332–340.

[5] P. B. Callahan and S. R. Kosaraju. A decomposition of multi-dimensional point-sets with applications to k-nearest-neighbors and n-body potential fields. *Proc. 24th ACM Symp. Theory of Computing* (1992) 546–556.

[6] K. L. Clarkson. Fast algorithms for the all-nearest-neighbors problem. *Proc. 24th IEEE Symp. Foundations of Computer Science* (1983) 226–232.

[7] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups*. Springer-Verlag. New York. 1988.

[8] L. Devroye. The expected size of some graphs in computational geometry. *Computers & Mathematics with Applications* 15 (1988) 53–64.

[9] M. T. Dickerson and D. Eppstein. Algorithms for proximity problems in higher dimensions. *Computational Geometry Theory & Applications* 5 (1996) 277–291.

[10] P. Eades and S. Whitesides. The realization problem for Euclidean minimum spanning trees is NP-hard. *Proc. 10th ACM Symp. Computational Geometry* (1994) 49–56.

[11] H. Edelsbrunner. *Algorithms in Combinatorial Geometry.* Springer-Verlag, 1987.

[12] D. Eppstein and J. Erickson. Iterated nearest neighbors and finding minimal polytopes. *Discrete & Computational Geometry* 11 (1994) 321–350.

[13] C. Monma and S. Suri. Transitions in geometric spanning trees. *Proc. 7th ACM Symp. Computational Geometry* (1991) 239–249.

[14] M. S. Paterson and F. F. Yao. On nearest-neighbor graphs. *Proc. 19th Int. Coll. Automata, Languages and Programming*, Springer LNCS 623 (1992) 416–426.

[15] S.-H. Teng and F. F. Yao. Percolation and k-nearest neighbor clustering. Manuscript, 1993.

[16] P. Vaidya. An $O(n \log n)$ algorithm for the all-nearest-neighbors problem. *Discrete & Computational Geometry* 4 (1989) 101–115.