A Lower Bound On Proximity Preservation by Space Filling Curves

Pan Xu
Industrial and Manufacturing Systems Engg.
Iowa State University
Ames, IA, USA
Email: panxu@iastate.edu

Srikanta Tirthapura
Electrical and Computer Engg.
Iowa State University
Ames, IA, USA
Email: snt@iastate.edu

***Abstract*—A space filling curve (SFC) is a proximity preserving mapping from a high dimensional space to a single dimensional space. SFCs have been used extensively in dealing with multi-dimensional data in parallel computing, scientific computing, and databases. The general goal of an SFC is that points that are close to each other in high-dimensional space are also close to each other in the single dimensional space. While SFCs have been used widely, the extent to which proximity can be preserved by an SFC is not precisely understood yet.**

We consider natural metrics, including the "nearest-neighbor stretch" of an SFC, which measure the extent to which an SFC preserves proximity. We first show a powerful negative result, that there is an inherent lower bound on the stretch of *any* SFC. We then show that the stretch of the commonly used Z curve is within a factor of 1.5 from the optimal, irrespective of the number of dimensions. Further we show that a very simple SFC also achieves the same stretch as the Z curve. Our results apply to SFCs in any dimension d such that d is a constant.

***Keywords*-space filling curve, proximity, stretch, lower bound**

I. Introduction

Space filling curves are a widely used tool for dealing with multi-dimensional data. The basic idea in a space filling curve is that data in two or more dimensions is mapped to a single dimension with the expectation that *proximity* is preserved. In other words, points that are close to each other in the high dimensional space are also (hopefully) close to each other in the space filling curve. Space filling curves have been used in numerous applications such as data partitioning in scientific computing [26, 23], parallel domain decomposition [3, 22], cryptography [16], secondary memory data structures [9], geographical information systems [1], to name a few.

Some popularly used space filling curves are the Z-curve [21, 19] (also known as the Morton ordering), the Hilbert curve [13], and the Gray code curve [9, 10]. While all the above applications of SFCs rely on the proximity preserving properties of SFCs, there has been surprisingly little formal analysis of the extent to which proximity is preserved by an SFC. There are two natural questions that one may ask in this context.

1) Are there any inherent limits on the extent to which proximity can be preserved by an SFC? If so, what are they?
2) How close are specific SFCs to the optimal SFC with respect to proximity preservation?

This work is an attempt to answer both these questions. In order to do so, we define precise metrics for evaluating the proximity preserving quality of an SFC.

Nearest Neighbor Stretch: We first consider proximity preservation for those pairs of points that are nearest neighbors in the multi-dimensional space. In many applications of SFCs, such as N-body simulations [26], the dominant interactions are the ones between nearest neighbors, and proximity preservation between such pairs of points is critical for the efficiency of the data structure.

To evaluate proximity preservation among nearest neighbors, we introduce a class of metrics called the "average nearest-neighbor stretch" of an SFC, henceforth referred to as the "average NN-stretch". Informally, the average NN-stretch of a d-dimensional SFC is the average multiplicative increase of the distance between nearest neighbor pairs in high-dimensions when the points are mapped into one dimension. We consider two variants of this definition, based on whether we consider the average distance to a nearest neighbor for each cell, followed by average across all cells (average-average NN-stretch), or the maximum distance to a nearest neighbor for each cell, followed by the average across all cells (average-maximum NN-stretch). In the following, we use "average NN-stretch" to mean the "average-average NN stretch". Precise definitions are provided in Section III. We show the following results.

- There is an inherent lower bound on the average NN-stretch of any SFC. In particular, the average NN-stretch of any SFC must be at least $\frac{2}{3d}\left(n^{1-\frac{1}{d}}\right)$ for large n, where n is the number of cells in the universe, and d is the number of dimensions (Theorem 1)
- The average NN-stretch of the Z space filling curve is within a factor of 1.5 of the above bound, irrespective

The work of Tirthapura was supported in part by the National Science Foundation through awards 0834743, 0831903

of the number of dimensions d (Theorem 2).
- Further, the average NN-stretch of a very simple space filling curve, which we call the “simple curve”, also has the same average NN-stretch as the Z curve (Theorem 3).

All Pairs Stretch: We next consider proximity preservation for all possible pairs of points, not just those pairs that are nearest neighbors in high dimensions. For this, we introduce a metric called the “average all pairs stretch”. Informally, the average all pairs stretch of a d dimensional SFC is the average multiplicative increase in the distance between a pair of points when the points are mapped from high dimensional space to one dimension. We consider two metrics for cells in high dimensions, the Manhattan metric and the Euclidean metric. Our results for the all pairs stretch include lower bounds for any SFC, as well as upper bounds for specific SFCs.

- For any SFC π, the average all pairs stretch using the Manhattan metric must be at least $\frac{1}{3d}\frac{n+1}{\sqrt[d]{n}-1} \approx \frac{n^{1-\frac{1}{d}}}{3d}$, where n is the size of the universe, and d is the number of dimensions.
- The average all pairs stretch of the simple curve is no more than $n^{1-\frac{1}{d}}$

In practice, proximity preservation is more important for pairs of points that are close to each other in high-dimensions than it is for points that are further apart in high-dimensional space. For example, if the application was to simulate N-body interactions between particles, the forces between particles get much weaker with distance, so that the interactions between particles that are far away are non-existent or negligible. Thus, we believe the average NN-stretch is usually a more significant metric than the average all pairs stretch.

Our results lead to the following observations:

1) For proximity preservation according to the average NN-stretch, the Z curve is close to optimal. To our knowledge, this is the first instance in the literature on space filling curves that we are able to make a precise statement about the optimality of an SFC with respect to proximity preservation. Previous works had investigated the properties of specific SFCs, but did not present a lower bound on the class of all SFCs.
2) Rather surprisingly, the simple curve has the same performance as the Z curve and is also near-optimal.
3) Further, a different space filling curve can yield only a constant factor improvement over the Z curve or the simple curve.

An SFC is generally thought of as a curve that does not intersect with itself. In this work, we define an SFC as a bijection from the high dimensional universe with n cells to the set $\{0, 1, 2, \ldots, n-1\}$. Since this can include curves that can self-intersect (see curve π_2 in Figure 1, for example), our definition of SFCs is a more general class than is usually considered. This also implies that each of our lower bounds is a lower bound for the class of non-intersecting SFCs also.

Roadmap: The rest of this paper is organized as follows. We discuss related work in Section II, followed by a discussion of the model and a definition of our metrics in Section III. We present our analysis of the NN-stretch in Section IV. This includes lower bounds for any SFC, as well as exact analysis of the Z curve and the simple curve. Related problems, including all pairs stretch, Euclidean metric, and variants are considered in Section V.

II. Related Work

There is a vast literature on SFCs. Here we attempt to cite closely related work. None of these works consider lower bounds for proximity preservation by an SFC, like we do.

Moon *et al.*[18] present a comprehensive analysis of the Hilbert SFC with respect to the “clustering” metric, defined as follows: given a rectangular region, into how many consecutive segments of an SFC can this be divided into, on average? The clustering metric is different from the metric that we consider. Our metric, the “stretch”, measures the extent to which distances are preserved. A related application of SFCs for managing multi-dimensional data in secondary memory is discussed in [9, 14].

Nearest-neighbor finding using SFCs has been studied in [5]. This work compares the Z-curve, the Gray-code curve, and the Hilbert-curve according to their performance for nearest-neighbor queries. They do not consider the stretch of an SFC, as we do here.

Neidermeier, Reinhardt, and Sanders [20] study the Hilbert SFC in two and three dimensions according to the ratio between distance between points on the two dimensional grid to the distance between them on the SFC. This work proves a result of the form: “If two points are indexed i and j on the Hilbert SFC, then their Manhattan distance on the two dimensional grid is bounded by $3\sqrt{i-j}-2$”. Thus, they prove that mapping from one dimension to two dimensions (usually) results in a contraction in distance, for the Hilbert curve. Note this result does not imply that mapping from two dimensions to one dimension results in a small expansion of distance. Indeed, as we show, there is an inherent lower bound on the extent to which this can be achieved. Thus, this work and our work are concerned with different metrics.

Dai and Su [7, 8] present upper bounds on the stretch of specific SFCs including the Hilbert curve, but do not consider lower bounds on the stretch, like we do here. Mitchison and Durbin [17] consider a metric close to the average NN-stretch as defined above, and present an analysis of the optimal numbering for the two dimensional case. When compared with this, our work presents an analysis of the problem for any (constant) number of dimensions. An even earlier work due to Harper [12] considers a metric that is related to the nearest neighbor stretch in high dimensions, but focuses on the n-cube, where the side length along each dimension is 2. In contrast, we consider a high-dimensional cube of side length 2^k.

Gotsman and Lindenbaum [11] also consider questions similar to [20]: to what extent can two points that are close to each other along the SFC be far apart in the multi-dimensional metric (Euclidean or Manhattan)? Similar to the above, our work considers a different metric, which goes the opposite direction: what is the extent to which distances are preserved when points are mapped from the multidimensional universe to the one dimensional universe.

Tirthapura, Seal and Aluru [25] considered the analysis of SFCs in the context of nearest neighbor queries and spherical region queries. They assume a probabilistic model of input and present an average case analysis of the performance of a class of SFCs that includes the Hilbert and Z curves. They do not consider the "stretch" as we define here, and restrict their analysis to two and three dimensions.

There is work on the efficient generation of the order of points according to the space filling curve. For example, SFCGen [15] proposes a table-driven approach to the generation of points along an SFC. Note that there have been several empirical comparisons of different space-filling curves, such as [1].

III. Model

The universe is the d dimensional grid of dimensions $\sqrt[d]{n} \times \cdots \times \sqrt[d]{n}$. For simplicity, we assume that $\sqrt[d]{n} = 2^k$ where k is a non-negative integer. Let U denote the set of all n cells. Each point in U is a d-tuple $(x_1, x_2, \ldots, x_d)$ where for each $i = 1 \ldots d$, $0 \leq x_i < \sqrt[d]{n}$.

For cells $\alpha, \beta \in U$, let $\Delta(\alpha, \beta)$ denote the Manhattan distance between α and β on the d dimensional grid. If $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_d)$, then $\Delta(\alpha, \beta) = \sum_{i=1}^{d} |\alpha_i - \beta_i|$.

A space filling curve (SFC) π is a bijection $\pi : U \rightarrow \{0, 1, \ldots, n-1\}$. A space filling curve provides a total order among all cells in U. For an SFC π, for cells $\alpha, \beta \in U$, let $\Delta_\pi(\alpha, \beta)$ denote $|\pi(\alpha) - \pi(\beta)|$, i.e., the distance between α and β on SFC π.

Nearest Neighbor Stretch: Our first metric captures the stretch between those pairs of cells that are nearest neighbors in U according to metric Δ. For cell $\alpha \in U$, let $N(\alpha)$ denote the set of cells $\beta \in U$ such that $\Delta(\alpha, \beta) = 1$. Note that cells that are nearest neighbors according to the Manhattan metric are also nearest neighbors according to the Euclidean metric in the d-dimensional space. The set $N(\alpha)$ is referred to as the "neighbors of α in U". It is clear that for each $\alpha \in U$, $d \leq |N(\alpha)| \leq 2d$. In the rest of the paper, we use the phrase "nearest neighbors" to mean the nearest neighbors according to the metric Δ, unless otherwise specified.

Definition 1. *For an SFC π and cell α, the average nearest neighbor stretch for α is defined as:*

$$\delta_\pi^{avg}(\alpha) = \frac{\sum_{\beta \in N(\alpha)} \Delta_\pi(\alpha, \beta)}{|N(\alpha)|}$$

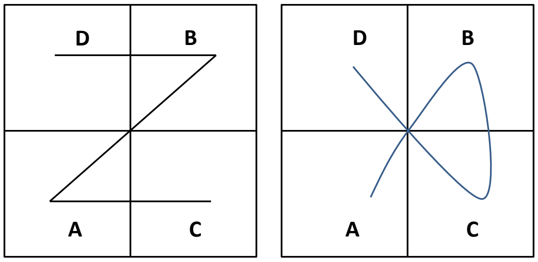


Figure 1. Two space filling curves on a 2×2 grid. The curves on the left and the right are referred to as π_1 and π_2 respectively. π_1 orders the cells as C, A, B, D and π_2 orders the cells as A, B, C, D.

Note that for any $\beta \in N(\alpha)$, it must be true that $\Delta(\alpha, \beta) = 1$, so the above expression can be interpreted as the average "dilation" of a path between nearest neighbors when the cells are organized using the SFC π. The intuition is that if the above metric is small, then the SFC preserves the distance between nearest neighbors well, and if the metric is large, then cells that are neighbors in U are far apart when organized using π.

Definition 2. *For a space filling curve π, the average-average nearest neighbor stretch for π is defined as:*

$$D^{avg}(\pi) = \frac{1}{n} \sum_{\alpha \in U} \delta_\pi^{avg}(\alpha)$$

We would like to have $D^{avg}(\pi)$ be as small as possible. For example, it would be best if $D^{avg}(\pi)$ was 1, so that cells that are nearest neighbors in U are assigned consecutive indices by π, but this is easily seen to be impossible. In subsequent sections, we present strong lower bounds for $D^{avg}(\pi)$, for a general SFC π.

Definition 3. *For an SFC π and cell α, the maximum nearest neighbor stretch for α is defined as:*

$$\delta_\pi^{max}(\alpha) = \max_{\beta \in N(\alpha)} \Delta_\pi(\alpha, \beta)$$

Definition 4. *The average-maximum nearest neighbor stretch for SFC π is defined as:*

$$D^{max}(\pi) = \frac{1}{n} \sum_{\alpha \in U} \delta_\pi^{max}(\alpha)$$

For the example shown in the Figure 1, we have the following. The values of $\delta^{avg}_{\pi_1}(A)$, $\delta^{avg}_{\pi_1}(B)$, $\delta^{avg}_{\pi_1}(C)$, and $\delta^{avg}_{\pi_1}(D)$ are all equal to 1.5, and hence $D^{avg}(\pi_1) = 1.5$. Similarly, it can be checked that $D^{avg}(\pi_2) = 2$, $D^{max}(\pi_1) = 2$, and $D^{max}(\pi_2) = 2.5$

We first prove some basic results about the distances induced by $\Delta_\pi(\cdot,\cdot)$.

Lemma 1. *For any space filling curve π on an universe U, Δ_π obeys the generalized triangle inequality. In other words, for any $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_k \in U$,*

$$\Delta_\pi(\alpha_1, \alpha_k) \leq \sum_{i=1}^{k-1} \Delta_\pi(\alpha_i, \alpha_{i+1})$$

Proof: We prove using mathematical induction. Consider the case when $k = 2$. Easily we can check that the generalized triangle inequality holds. Now assume the generalized triangle inequality holds for each $k \leq \ell - 1$, then we have:

$$\begin{aligned}\Delta_\pi(\alpha_1, \alpha_\ell) &\leq \Delta_\pi(\alpha_1, \alpha_{\ell-1}) + \Delta_\pi(\alpha_{\ell-1}, \alpha_\ell) \\ &\leq \sum_{i=1}^{n-2} \Delta_\pi(\alpha_i, \alpha_{i+1}) + \Delta_\pi(\alpha_{\ell-1}, \alpha_\ell) \\ &= \sum_{i=1}^{\ell-1} \Delta_\pi(\alpha_i, \alpha_{i+1})\end{aligned}$$

So we have that the generalized triangle inequality holds for $k = n$.

■

IV. Analysis of Average Nearest Neighbor Stretch

We first present a lower bound on $D^{avg}(\pi)$ for any SFC π. Then, we present an exact analysis of $D^{avg}(Z)$ where Z is the Z-curve on d dimensions. We follow this by the exact analysis of a space filling curve S with a simple structure, and show that $D^{avg}(S)$ matches that of the Z curve.

A. Lower Bound

In this section, we present a lower bound for the average-average NN-stretch for any SFC. Let A be the set of all possible unordered pairs in U while A' be the set of all possible ordered pairs in U.

Let NN^d denote the set of unordered pairs that are nearest neighbors in the d dimensional universe, i.e.

$$NN^d = \{(\alpha, \beta) | \alpha \in U, \beta \in U, \Delta(\alpha, \beta) = 1\}$$

Think of elements of NN^d as the "edges" of length 1 between cells in the metric defined by Δ.

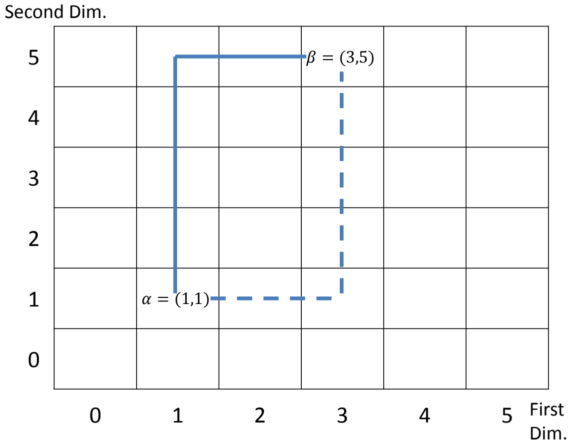


Figure 2. The dashed path denotes $p(\alpha, \beta)$, and the solid path denotes $p(\beta, \alpha)$.

Theorem 1. *For any SFC π whose domain is a d-dimensional universe U, it must be true that:*

$$D^{avg}(\pi) \geq \frac{2}{3d}(n^{1-\frac{1}{d}} - n^{-1-\frac{1}{d}})$$

To prove the above theorem, we introduce the notion of a "nearest neighbor decomposition" of a pair of cells $(\alpha, \beta) \in A'$, denoted by $p(\alpha, \beta)$, defined as follows. Let the coordinates of the cells be $\alpha = (x_1, \cdots, x_d), \beta = (y_1, \cdots, y_d)$. Intuitively, $p(\alpha, \beta)$ defines a specific path from α to β, and can thus be considered as a set of edges (unordered pairs of vertices), i.e. a subset of NN^d.

Suppose the coordinates of α and β differed only along one dimension, say $i, 1 \leq i \leq d$. Then when $x_i < y_i$, we have:

$$p(\alpha, \beta) = \bigcup_{\ell=x_i}^{y_i-1} \{((x_1, \ldots, \ell, \ldots, x_d), (x_1, \ldots, \ell{+}1, \ldots, x_d))\}$$

and when $x_i > y_i$, we have:

$$p(\alpha, \beta) = \bigcup_{\ell=y_i}^{x_i-1} \{((x_1, \ldots, \ell, \ldots, x_d), (x_1, \ldots, \ell{+}1, \ldots, x_d))\}$$

For example, $p((6,4,5),(3,4,5))$ is the set $\{((3,4,5),(4,4,5)),((4,4,5),(5,4,5)),((5,4,5),(6,4,5))\}$. Note that if α and β differed only in one coordinate, $p(\alpha, \beta)$ equals $p(\beta, \alpha)$.

If the coordinates of α and β differ along more than one dimension, then we define the following sequence of vertices that form a path from α to β. In this path, we "correct" the coordinates of α one dimension at a time, starting from dimension 1 till dimension d, until we reach β. $[\alpha_0 = \alpha = (x_1, \cdots, x_d)], [\alpha_1 = (y_1, x_2, \cdots, x_d)], [\alpha_2 = (y_1, y_2, x_3, \cdots, x_d)], \cdots [\alpha_{d-1} =$

$(y_1, \cdots, y_{d-1}, x_d)], [\alpha_d = \beta = (y_1, \cdots, y_{d-1}, y_d)]$. Note that α_i and α_{i+1} differ only along a single dimension. Then we have:

$$p(\alpha, \beta) = \bigcup_{i=0}^{d-1} p(\alpha_i, \alpha_{i+1})$$

Note that for a pair of cells $(\alpha, \beta) \in A'$, $p(\alpha, \beta)$ can be different from $p(\beta, \alpha)$. In Figure 2, $p(\alpha, \beta)$ is the set $\{((1,1),(2,1)), ((2,1),(3,1)), ((3,1),(3,2)), ((3,2),(3,3)), ((3,3),(3,4)), ((3,4),(3,5))\}$ while $p(\beta, \alpha)$ is the set $\{((1,5),(2,5)), ((2,5),(3,5)), ((1,1),(1,2)), ((1,2),(1,3)), ((1,3),(1,4)), ((1,4),(1,5))\}$

Our strategy for a proof of Theorem 1 is as follows: We compute $\sum_{(\alpha,\beta)\in A'} \Delta_\pi(\alpha, \beta)$ in two different ways. In one manner, we compute it directly and exactly, leading to a bound of $\Theta(n^3)$. In another way, we find an upper bound in terms of $D^{avg}(\pi)$, using the nearest neighbor decomposition of each $(\alpha, \beta) \in A'$, and then the triangle inequality. This eventually leads to a lower bound on $D^{avg}(\pi)$. Let $S_{A'}(\pi)$ be defined as follows.

$$S_{A'}(\pi) = \sum_{(\alpha,\beta)\in A'} \Delta_\pi(\alpha, \beta)$$

Lemma 2. *For any SFC π,*

$$S_{A'}(\pi) = \frac{1}{3}(n-1)n(n+1) \tag{1}$$

Proof: We partition A' into $n-1$ subgroups by the distance on π. Let $A'_i \subseteq A', 1 \le i \le n-1$ denote the group of pairs such that $\Delta_\pi(\alpha, \beta) = i, \forall (\alpha, \beta) \in A'_i$. Then easily we get that $|A'_i| = 2(n-i)$. So we have:

$$S_{A'}(\pi) = \sum_{i=1}^{n-1} 2i(n-i) = \frac{1}{3}(n-1)n(n+1)$$

■

Lemma 3. *For any d dimensional SFC π, we have:*

$$\begin{aligned} \frac{1}{nd} \sum_{(\alpha,\beta)\in NN^d} \Delta_\pi(\alpha, \beta) &\le D^{avg}(\pi) \\ D^{avg}(\pi) &\le \frac{2}{nd} \sum_{(\alpha,\beta)\in NN^d} \Delta_\pi(\alpha, \beta) \end{aligned}$$

Proof: From the definition $D^{avg}(\pi)$, we have:

$$\begin{aligned} D^{avg}(\pi) &= \frac{1}{n} \sum_{\alpha\in U} \frac{1}{|N(\alpha)|} \sum_{\beta\in N(\alpha)} \Delta_\pi(\alpha, \beta) \\ &= \frac{1}{n} \sum_{(\alpha,\beta)\in NN^d} \left(\frac{1}{|N(\alpha)|} + \frac{1}{|N(\beta)|} \right) \Delta_\pi(\alpha, \beta) \end{aligned}$$

Note that for each cell $\alpha \in U$, $d \le |N(\alpha)| \le 2d$. So we have $\frac{1}{d} \le \frac{1}{|N(\alpha)|} + \frac{1}{|N(\beta)|} \le \frac{2}{d}$. Combining these two inequalities with the equality above yields our conclusion.

■

Now we are ready to prove Theorem 1.

Proof of Theorem 1: For every pair $(\alpha, \beta) \in A'$, note that the vertex pairs in $p(\alpha, \beta)$ together form a path from α to β. Thus, we have the following from Lemma 1 (generalized triangle inequality):

$$\Delta_\pi(\alpha, \beta) \le \sum_{(\alpha',\beta')\in p(\alpha,\beta)} \Delta_\pi(\alpha', \beta') \tag{2}$$

It follows that:

$$\sum_{(\alpha,\beta)\in A'} \Delta_\pi(\alpha, \beta) \le \sum_{(\alpha,\beta)\in A'} \sum_{(\alpha',\beta')\in p(\alpha,\beta)} \Delta_\pi(\alpha', \beta') \tag{3}$$

We show in Lemma 4 that for each neighboring pair $(\alpha', \beta') \in NN^d$, it appears in the right side of inequality 3 at most $\frac{1}{2}n^{\frac{d+1}{d}}$ times. So we have the following inequality:

$$\sum_{(\alpha,\beta)\in A'} \Delta_\pi(\alpha, \beta) \le \frac{1}{2} n^{\frac{d+1}{d}} \sum_{(\zeta,\eta)\in NN^d} \Delta_\pi(\zeta, \eta) \tag{4}$$

Recall that $\sum_{(\alpha,\beta)\in A'} \Delta_\pi(\alpha, \beta) = \frac{1}{3}(n^3 - n)$ from Lemma 2 and $D^{avg}(\pi) \ge \frac{1}{nd} \sum_{(\alpha,\beta)\in NN^d} \Delta_\pi(\alpha, \beta)$ from Lemma 3. So after combining Lemma 2, Lemma 3 and inequality 4 we get $D^{avg}(\pi) \ge \frac{2}{3}\frac{1}{d}(n^{1-\frac{1}{d}} - n^{-1-\frac{1}{d}})$. ■

Lemma 4. *For each neighboring pair $(\zeta, \eta) \in NN^d$, there exist at most $\frac{1}{2}n^{\frac{d+1}{d}}$ pairs (α, β) in A' such that $(\zeta, \eta) \in p(\alpha, \beta)$.*

Proof: Assume $\zeta = (\zeta_1, \cdots, \zeta_i, \cdots, \zeta_d), \eta = (\zeta_1, \cdots, \zeta_i + 1, \cdots, \zeta_d), 1 \le i \le d$, i.e. (ζ, η) differs in the i coordinate.

Let $\alpha = (x_1, \cdots, x_d), \beta = (y_1, \cdots, y_d)$. According to our nearest neighbor decomposition, we have:

$$p(\alpha, \beta) = \bigcup_{j=0}^{d-1} p(\alpha_j, \alpha_{j+1})$$

Note that $p(\alpha_j, \alpha_{j+1}), 0 \le j \le d-1$ consists of all neighboring pairs which differs in the $(j+1)$ coordinate. So easily we have that:

$$(\zeta, \eta) \in p(\alpha, \beta) \Longleftrightarrow (\zeta, \eta) \in p(\alpha_{i-1}, \alpha_i)$$

Note that when $x_i < y_i$, we have $p(\alpha_{i-1}, \alpha_i) =$

$$\bigcup_{\ell=x_i}^{y_i-1} \{((y_1, \ldots, y_{i-1}, \ell, x_{i+1}, \ldots, x_d), (y_1, \ldots, y_{i-1}, \ell+1, x_{i+1}, \ldots, x_d))\}$$

while when $x_i > y_i$, we have $p(\alpha_{i-1}, \alpha_i) =$

$$\bigcup_{\ell=y_i}^{x_i-1} \{((y_1, \ldots, y_{i-1}, \ell, x_{i+1}, \ldots, x_d), \\ (y_1, \ldots, y_{i-1}, \ell + 1, x_{i+1}, \ldots, x_d))\}$$

It follows that $(\zeta, \eta) \in p(\alpha_{i-1}, \alpha_i)$ if and only if (1) the first $i-1$ coordinates of β must share with the first $i-1$ coordinates of ζ; (2) the last $d-i$ coordinates of α must share with the last $d-i$ coordinates of ζ; (3) the interval between ζ_i and $\zeta_i + 1$ must be contained in the interval between x_i and y_i. The exact mathematical description is as follows:

$$(\zeta, \eta) \in p(\alpha_{i-1}, \alpha_i) \Longleftrightarrow \begin{cases} y_j = \zeta_j, 1 \leq j \leq i-1 \\ x_j = \zeta_j, i+1 \leq j \leq d \\ (x_i \leq \zeta_i < y_i) \vee (y_i \leq \zeta_i < x_i) \end{cases}$$

So the total number of possible pairs (α, β) such that $(\zeta, \eta) \in p(\alpha, \beta)$ should be $2(\sqrt[d]{n})^{d-1}\zeta_i(\sqrt[d]{n} - \zeta_i)$. Easily we can get the total number is upper bounded by $\frac{1}{2}n^{\frac{d+1}{d}}$. ■

B. Performance of Z Curve

In this section we will consider the d dimensional Z curve (see Figure 3). We show that the average-average NN-stretch of the Z curve is within a factor of 1.5 of the lower bound, and hence within a factor of 1.5 of the optimal. Recall that $\sqrt[d]{n} = 2^k$. A cell x in the universe U can be represented by coordinates $(x_1, x_2, \ldots, x_d)$, where for $1 \leq i \leq d$, $0 \leq x_i < 2^k$; thus x_i can be represented using a binary string of length k. For $j = 1 \ldots d$, let x_i^j denote the j-th most significant bit in x_i.

We recall the definition of a d dimensional Z curve ([21, 19]). The Z curve is defined by assigning to each cell $x \in U$, a "key" $Z(x)$, which is an integer that denotes the position of a cell in the space filling curve order. $Z(x)$ *is equal to the binary number represented by the string* $x_1^1, x_2^1, \cdots, x_d^1, x_1^2, x_2^2, \cdots, x_d^2, \cdots, x_1^k, x_2^k, \cdots, x_d^k$. In other words, the coordinates in different dimensions are interleaved together to form the key of the cell. For example, if $d = 3, k = 3$, then $Z(101, 010, 011) = 100011101$. Note that different Z curves are possible by taking the dimensions in a different order during interleaving, but these are all equivalent to the above definition, at least for the metrics that we consider.

Notation: We write $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$.

Theorem 2. *If Z is the d dimensional Z-curve, then*

$$D^{avg}(Z) \sim \frac{1}{d} n^{1-\frac{1}{d}}$$

Our strategy for a proof of the above theorem is as follows. We divide NN^d into d groups $\{G_i\}, 1 \leq i \leq d$ as follows. G_i is the set of neighboring pairs $(\alpha, \beta) \in NN^d$ such that α and β differ in the ith coordinate. Clearly the different G_i are all disjoint. Let $\Lambda_i(Z)$ denotes the total sum of distances on Z curve for all pairs in G_i, i.e.

$$\Lambda_i(Z) = \sum_{(\alpha,\beta)\in G_i} \Delta_Z(\alpha, \beta)$$

Lemma 5.

$$\lim_{n\to\infty} \frac{\Lambda_i(Z)}{n^{2-\frac{1}{d}}} = \frac{2^{d-i}}{2^d - 1}, \forall 1 \leq i \leq d$$

Proof: Consider a neighboring pair $(\alpha, \beta) \in G_i$. These two differ only in dimension i, and suppose that the coordinates of α and β in dimension i are κ and $\kappa + 1$ respectively.

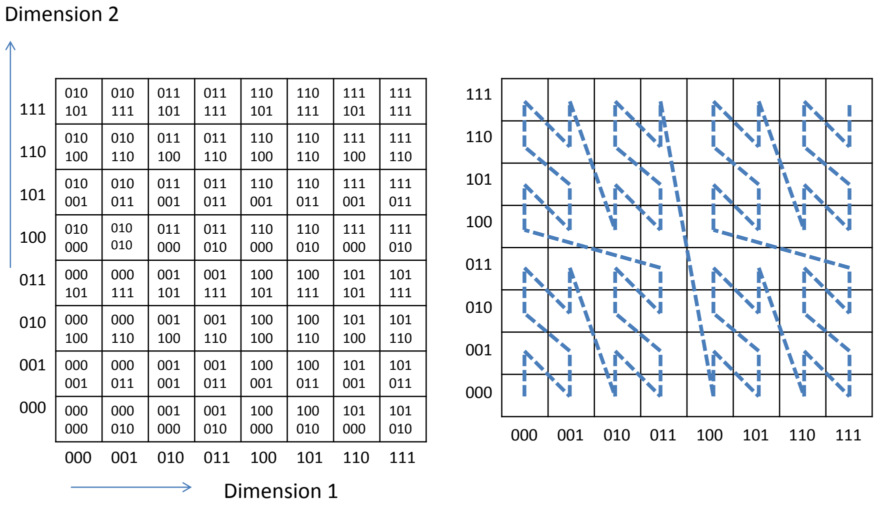
First consider the case when the least significant bit of κ is 0. In this case the coordinates of α and β in dimension i differ only in the least significant bit, and their coordinates are equal in all other dimensions. By the definition of the Z curve, $\Delta_Z(\alpha, \beta) = |Z(\alpha) - Z(\beta)| = 2^{d-i}$. The total number of pairs in G_i where the least significant bit of κ is 0 is $2^{k-1}n^{1-\frac{1}{d}}$. This is because the ith coordinate of α can be chosen in 2^{k-1} ways and the other coordinates of α can be chosen in $n^{1-\frac{1}{d}}$ ways, and these choices are independent of each other. Once α is chosen, β is also fixed.

For $j = 1 \ldots k$, let $G_{i,j} \subseteq G_i$ denote the subset of pairs $(\alpha, \beta) \in G_i$ such that the $j-1$ least significant bits of κ are 1, and the jth least significant of κ is 0. i.e. κ has the form $(*, *, \ldots, 0, 1, 1, (j-1) \quad \text{times}, 1)$. In this case, $\kappa+1$ has the form $(*, *, \ldots, 1, 0, 0, (j-1) \quad \text{times}, 0)$. By the definition of the Z curve, we get for all $(\alpha, \beta) \in G_{i,j}$:

$$\Delta_Z(\alpha, \beta) = 2^{jd-i} - \sum_{\ell=1}^{j-1} 2^{\ell d - i}$$

The total number of neighboring pairs in $G_{i,j}$ is $2^{k-j}n^{1-\frac{1}{d}}$.

$$\begin{aligned} \Lambda_i(Z) &= \sum_{j=1}^{k} |G_{i,j}| \left(2^{jd-i} - \sum_{\ell=1}^{j-1} 2^{\ell d-i}\right) \\ &= \sum_{j=1}^{k} 2^{k-j} n^{1-\frac{1}{d}} \left(2^{jd-i} - \sum_{\ell=1}^{j-1} 2^{\ell d-i}\right) \\ &= \sum_{j=1}^{k} 2^{k-j} n^{1-\frac{1}{d}} \left(2^{jd-i} - \frac{2^{jd-i} - 2^{d-i}}{2^d - 1}\right) \\ &= \sum_{j=1}^{k} 2^{k-j} n^{1-\frac{1}{d}} \left(\frac{2^d - 2}{2^d - 1} 2^{jd-i} + \frac{2^{d-i}}{2^d - 1}\right). \end{aligned}$$



111	010 101	010 111	011 101	011 111	110 101	110 111	111 101	111 111
110	010 100	010 110	011 100	011 110	110 100	110 110	111 100	111 110
101	010 001	010 011	011 001	011 011	110 001	110 011	111 001	111 011
100	010 000	010 010	011 000	011 010	110 000	110 010	111 000	111 010
011	000 101	000 111	001 101	001 111	100 101	100 111	101 101	101 111
010	000 100	000 110	001 100	001 110	100 100	100 110	101 100	101 110
001	000 001	000 011	001 001	001 011	100 001	100 011	101 001	101 011
000	000 000	000 010	001 000	001 010	100 000	100 010	101 000	101 010
	000	001	010	011	100	101	110	111

Figure 3. Two dimensional Z curve on an 8×8 grid. On the left is the assignment of keys to cells; within each cell the higher order bits are on the top and the lower order bits are at the bottom. On the right is a pictorial representation of the order.

Let

$$\begin{aligned} Q_1 &= \sum_{j=1}^{k} 2^{k-j} n^{1-\frac{1}{d}} \left(\frac{2^d-2}{2^d-1} 2^{jd-i} \right) \\ Q_2 &= \sum_{j=1}^{k} 2^{k-j} n^{1-\frac{1}{d}} \left(\frac{2^{d-i}}{2^d-1} \right). \end{aligned}$$

Recall that $n = 2^{kd}$. Thus for Q_2, we have:

$$\begin{aligned} \lim_{n\to\infty} \frac{Q_2}{n^{2-\frac{1}{d}}} &= \lim_{n\to\infty} \frac{2^{d-i}}{2^d-1} \frac{2^k-1}{n} \\ &= 0 \end{aligned}$$

For Q_1, we have:

$$\begin{aligned} \lim_{n\to\infty} \frac{Q_1}{n^{2-\frac{1}{d}}} &= \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{k} 2^{k-i+j(d-1)} \left(\frac{2^d-2}{2^d-1} \right) \\ &= \frac{2^{d-i}}{2^d-1} \end{aligned}$$

Finally, we get:

$$\lim_{n\to\infty} \frac{\Lambda_i(Z)}{n^{2-\frac{1}{d}}} = \frac{2^{d-i}}{2^d-1}$$

■

Now we are ready to prove Theorem 2.

Proof of Theorem 2:

We partition the set NN^d into two subsets.

Recall that

$$D^{avg}(Z) = \frac{1}{n} \sum_{(\alpha,\beta)\in NN^d} \left(\frac{1}{|N(\alpha)|} + \frac{1}{|N(\beta)|} \right) \Delta_Z(\alpha, \beta)$$

Let H_1 denote the set of pairs $\{(\alpha, \beta) \in NN^d | N(\alpha) = N(\beta) = 2d\}$

Let H_2 denote the set of pairs $\{(\alpha, \beta) \in NN^d | (N(\alpha) < 2d) \vee (N(\beta) < 2d)\}$

We have

$$D^{avg}(Z) = \frac{1}{n}(h_1 + h_2)$$

where

$$h_1 = \frac{1}{d} \sum_{(\alpha,\beta)\in NN^d} \Delta_Z(\alpha, \beta) \tag{5}$$

and

$$h_2 = \sum_{(\alpha,\beta)\in H_2} \left(\frac{1}{|N(\alpha)|} + \frac{1}{|N(\beta)|} - \frac{1}{d} \right) \Delta_Z(\alpha, \beta) \tag{6}$$

We first evaluate h_1. By the definition of Λ_i, we have:

$$h_1 = \frac{1}{d} \sum_{i=1}^{d} \Lambda_i(Z)$$

Using Lemma 5, we get:

$$\lim_{n\to\infty} \frac{h_1}{n^{2-\frac{1}{d}}} = \frac{1}{d} \sum_{i=1}^{d} \frac{2^{d-i}}{2^d-1} = \frac{1}{d} \tag{7}$$

We now consider h_2. We know for any $\alpha \in U$, $|N(\alpha)| \geq$

d. Using this in Equation 6 leads to:

$$h_2 \leq \frac{1}{d} \sum_{(\alpha,\beta)\in H_2} \Delta_Z(\alpha,\beta)$$

Note that in every pair $(\alpha,\beta) \in H_2$, it must be true that for some $i, 1 \leq i \leq d$, either (1)the ith coordinate of α is 0 or $2^k - 1$, or (2)the ith coordinate of β is 0 or $2^k - 1$. Suppose for each neighboring pair $(\alpha,\beta) \in NN^d$, we use $(\alpha,\beta)_i$ to denote the pair of the ith coordinate of α and β. Let $K_1 = \{(\alpha,\beta) \in NN^d | \exists 1 \leq i \leq d : (\alpha,\beta)_i = (0,1) \vee (\alpha,\beta)_i = (2^k-2, 2^k-1)\}$, $K_2 = \{(\alpha,\beta) \in NN^d | \exists 1 \leq i \leq d : (\alpha,\beta)_i = (0,0) \vee (\alpha,\beta)_i = (2^k-1, 2^k-1)\}$. Then we have $H_2 = K_1 \bigcup K_2$ while K_1 and K_2 are disjoint.

Let $\Lambda(K_1)$, $\Lambda(K_2)$ and $\Lambda(H_2)$ denote the total sum of distances on Z curve for all pairs in $\Lambda(K_1)$, $\Lambda(K_2)$ and $\Lambda(H_2)$ respectively. We first analyze $\Lambda(K_1)$ as follows.

Let $K_{1,i} = \{(\alpha,\beta) \in NN^d | (\alpha,\beta)_i = (0,1) \vee (\alpha,\beta)_i = (2^k-2, 2^k-1)\}$. So we have $K_1 = \bigcup_{i=1}^d K_{1,i}$. Note that the least significant bits of 0 and $2^k - 2$ are both 0. From the definition of $G_{i,1}$ we have $K_{1,i} \subseteq G_{i,1}, \forall 1 \leq i \leq d$. So we have $\Delta_Z(\alpha,\beta) = 2^{d-i}, \forall(\alpha,\beta) \in K_{1,i}, \forall 1 \leq i \leq d$. Easily we can get that the total number of pairs in $K_{1,i}$ should be $2n^{1-\frac{1}{d}}$. So we have:

$$\begin{aligned}\Lambda(K_1) &= \sum_{i=1}^{d} \sum_{(\alpha,\beta)\in K_{1,i}} \Delta_Z(\alpha,\beta) \\ &= \sum_{i=1}^{d} 2n^{1-\frac{1}{d}} 2^{d-i} \\ &= 2n^{1-\frac{1}{d}}(2^d - 1)\end{aligned}$$

Now we turn to $\Lambda(K_2)$. Let $K_{2,i} = \{(\alpha,\beta) \in NN^d | (\alpha,\beta)_i = (0,0) \vee (\alpha,\beta)_i = (2^k-1, 2^k-1)\}$. We can analyze $K_{2,i}$ in the same way as we analyze NN^d. What is different here is one of coordinates in the pair is fixed at 0 or $2^k - 1$. So we have:

$$\Lambda(K_{2,i}) = \sum_{\ell=1,\ell\neq i}^{d} 2\frac{\Lambda_\ell(Z)}{n^{\frac{1}{d}}}$$

It follows that:

$$\begin{aligned}\lim_{n\to\infty} \frac{\Lambda(K_2)}{n^{2-\frac{1}{d}}} &\leq \lim_{n\to\infty} \frac{\sum_{i=1}^{d} \Lambda(K_{2,i})}{n^{2-\frac{1}{d}}} \\ &\leq \lim_{n\to\infty} 2n^{-\frac{1}{d}} \frac{\sum_{i=1}^{d}\sum_{\ell=1}^{d} \Lambda_\ell(Z)}{n^{2-\frac{1}{d}}} \\ &= 0\end{aligned}$$

Easily we can get that:

$$\lim_{n\to\infty} \frac{\Lambda(K_1)}{n^{2-\frac{1}{d}}} = 0$$

So we have:

$$\lim_{n\to\infty} \frac{\Lambda(H_2)}{n^{2-\frac{1}{d}}} = \lim_{n\to\infty} \frac{\Lambda(K_1)+\Lambda(K_2)}{n^{2-\frac{1}{d}}} = 0$$

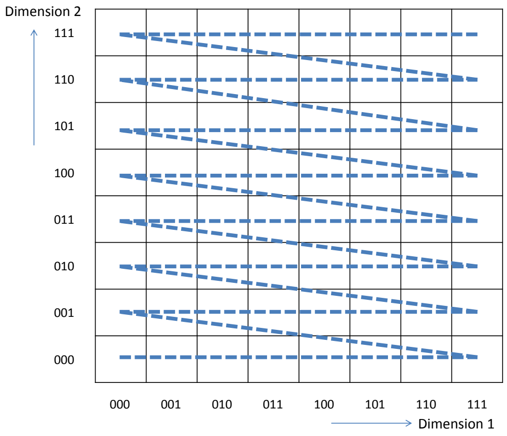


Figure 4. A simple curve S on an 8×8 grid.

Now we have:

$$\begin{aligned}\lim_{n\to\infty} \frac{h_2}{n^{2-\frac{1}{d}}} &\leq \frac{1}{d} \lim_{n\to\infty} \frac{\Lambda(H_2)}{n^{2-\frac{1}{d}}} \\ &= 0\end{aligned}$$

So we have $\lim_{n\to\infty} \frac{h_2}{n^{2-\frac{1}{d}}} = 0$. Combining the limitation of h_1 and h_2 yields:

$$\lim_{n\to\infty} \frac{h_1 + h_2}{n^{2-\frac{1}{d}}} = \frac{1}{d}$$

Note that $D^{avg}(Z) = \frac{1}{n}(h_1 + h_2)$. Thus we get $D^{avg}(Z) \sim \frac{1}{d} n^{1-\frac{1}{d}}$.

■

C. Performance of a Simple Space-Filling Curve

In this section we will show that even a simple curve, the one shown in Figure 4 can have the same performance as the Z-curve.

First we discuss how to construct d dimensional simple curve filling the grid of size $\sqrt[d]{n} \times \cdots \times \sqrt[d]{n}$. Refer to each cell as a tuple of coordinate $(x_1, \ldots, x_d)$, where for $1 \leq i \leq d, 0 \leq x_i < \sqrt[d]{n}$. For each cell $\alpha = (x_1, \ldots, x_d)$, the linear order assigned by simple curve $S(\alpha)$ can be defined as follows:

$$S(\alpha) = \sum_{i=1}^{d} x_i (\sqrt[d]{n})^{i-1} \qquad (8)$$

In the rest sections, once we mention d dimensional simple curve S, we mean it refers to the specific simple curve defined in Equation 8.

Theorem 3. *If S is the d dimensional simple curve, then*

$$D^{avg}(S) \sim \frac{1}{d} n^{1-\frac{1}{d}}$$

Proof: Recall that U denote the set of all n cells. Set $U_1 = \{(x_1, \ldots, x_d) : 1 \le x_i \le \sqrt[d]{n} - 2, \forall 1 \le i \le d\}, U_2 = \{(x_1, \ldots, x_d) : \exists 1 \le i \le d, x_i = 0 \text{ or } \sqrt[d]{n} - 1\}$. Here U_1 refers to the subset of cells forming a d dimensional sub grid while U_2 refers to the subset of cells forming the $2d$ $(d-1)$-dimensional faces. Obviously, $U = U_1 \cup U_2$.

For each cell $(x_1, \ldots, x_d)$ in U_1, it definitely has $2d$ nearest neighbors since $1 \le x_i \le \sqrt[d]{n} - 2, \forall 1 \le i \le d$. The total number of cells in U_1 is $(\sqrt[d]{n} - 2)^d$. For each cell $\alpha \in U_1$, the average nearest neighbor stretch $\delta_S^{avg}(\alpha) = \frac{1}{d} \sum_{\ell=0}^{d-1} (\sqrt[d]{n})^\ell = \frac{1}{d} \frac{n-1}{\sqrt[d]{n}-1}$. Denote the total sum of average nearest neighbor stretch for all cells in U_1 as $\Lambda(U_1)$, i.e, $\Lambda(U_1) = \sum_{\alpha \in U_1} \delta_S^{avg}(\alpha)$. Then we have:

$$\begin{aligned} \lim_{n \to \infty} \frac{\Lambda(U_1)}{n^{2-\frac{1}{d}}} &= \lim_{n \to \infty} \frac{1}{d} \frac{n-1}{\sqrt[d]{n}-1} (\sqrt[d]{n} - 2)^d \frac{1}{n^{2-\frac{1}{d}}} \\ &= \frac{1}{d} \end{aligned}$$

The total number of cells in U_2 should be $n - (\sqrt[d]{n} - 2)^d$. For each cell α in U_2, we have the average nearest neighbor stretch $\delta_S^{avg}(\alpha) \le \frac{2}{d} \frac{n-1}{\sqrt[d]{n}-1}$. Denote the total sum of average nearest neighbor stretch for all cells in U_2 as $\Lambda(U_2)$, i.e, $\Lambda(U_2) = \sum_{\alpha \in U_2} \delta_S^{avg}(\alpha)$. Then we have:

$$\begin{aligned} \lim_{n \to \infty} \frac{\Lambda(U_2)}{n^{2-\frac{1}{d}}} &\le \lim_{n \to \infty} \frac{2}{d} \frac{n-1}{\sqrt[d]{n}-1} \left(n - (\sqrt[d]{n} - 2)^d\right) \frac{1}{n^{2-\frac{1}{d}}} \\ &= 0 \end{aligned}$$

So we get that $\lim_{n \to \infty} \frac{\Lambda(U_2)}{n^{2-\frac{1}{d}}} = 0$. Combining the two together we have:

$$\lim_{n \to \infty} \frac{1}{n^{2-\frac{1}{d}}} (\Lambda(U_1) + \Lambda(U_2)) = \frac{1}{d}$$

Note that the average-average nearest neighbor stretch for all cells in U should be $D^{avg}(S) = \frac{1}{n}(\Lambda(U_1) + \Lambda(U_2))$. Thus we have $D^{avg}(S) \sim \frac{1}{d} n^{1-\frac{1}{d}}$.

■

V. Extensions

In this section, we present results for other variants of the definitions of the stretch. First, we consider the average-maximum NN stretch D^{max}. Then we consider the average stretch between all pairs of cells, not just those pairs that are nearest neighbors. We consider two variants of this "all pairs stretch", one where the metric in high-dimensions is the Manhattan metric, and the other where the metric in higher dimensions is the Euclidean metric. For each of these problems, we present upper and lower bounds, to the extent possible.

A. Maximum Nearest Neighbor Stretch

Recall that the average-maximum nearest neighbor stretch of a cell α is defined as:

$$\delta_\pi^{max}(\alpha) = \max_{\beta \in N(\alpha)} \Delta_\pi(\alpha, \beta)$$

The average-maximum NN-stretch of SFC π is: $D^{max}(\pi) = \frac{1}{n} \sum_{\alpha \in U} \delta_\pi^{max}(\alpha)$. We have the following lower bound.

Proposition 1. *For any SFC π whose domain is a d-dimensional universe U, it must be true that:*

$$D^{max}(\pi) \ge \frac{2}{3d} (n^{1-\frac{1}{d}} - n^{-1-\frac{1}{d}})$$

Proof: The inequality can be easily obtained from the following fact that for each $\alpha \in U$,

$$\delta_\pi^{max}(\alpha) \ge \delta_\pi^{avg}(\alpha)$$

.

It follows that $D^{max}(\pi) \ge D^{avg}(\pi)$. Combining the result in Theorem 1 yields the inequality above. ■

In the following, we show that the lower bound obtained above is nontrivial by showing the performance of the simple curve defined in 8 can be $n^{1-\frac{1}{d}}$, i.e, the simple curve is optimal up to a factor equal to the number of dimensions d.

Proposition 2. *If S is the d dimensional simple curve, then*

$$D^{max}(S) = n^{1-\frac{1}{d}}$$

Proof: Recall that for each cell $\alpha = (x_1, \ldots, x_d)$, the linear order assigned by simple curve $S(\alpha)$ is defined as follows:

$$S(\alpha) = \sum_{i=1}^{d} x_i (\sqrt[d]{n})^{i-1} \qquad (9)$$

For each cell $\alpha = (x_1, \ldots, x_d)$, at least one of the two neighbors, say $\beta_1 = (x_1, \ldots, x_d + 1)$ and $\beta_1 = (x_1, \ldots, x_d - 1)$, should exist. Note that $\Delta_S(\alpha, \beta_1) = \Delta_S(\alpha, \beta_2) = n^{1-\frac{1}{d}}$.

So we have for each cell $\alpha \in U$:

$$\delta_S^{max}(\alpha) = n^{1-\frac{1}{d}}$$

It follows that $D^{max}(S) = n^{1-\frac{1}{d}}$. ■

The above result should be compared with the average-average NN-stretch of the simple curve (Theorem 3). This shows that the average-maximum stretch is worse that the average-average stretch by a factor d. The intuitive explanation for this is that for a vast majority of cells that do not lie on the border, the distance (along the SFC) to two of the nearest neighbors is large, while the other $2d - 2$ nearest neighbors are much closer along the SFC.

B. All Pairs Stretch

In the case of all pairs stretch, the distance between two cells α and β in high dimensions can be different depending on whether we use the Euclidean or the Manhattan metric.

The average all pairs stretch for π, using the Manhattan metric is:

$$str^{avg,M}(\pi) = \frac{2}{n(n-1)} \sum_{(\alpha,\beta) \in A} \frac{\Delta_\pi(\alpha, \beta)}{\Delta(\alpha, \beta)}$$

Let Δ_E denote the Euclidean metric on universe U. The average all pairs stretch for π, using the Euclidean metric is:

$$str^{avg,E}(\pi) = \frac{2}{n(n-1)} \sum_{(\alpha,\beta)\in A} \frac{\Delta_\pi(\alpha,\beta)}{\Delta_E(\alpha,\beta)}$$

In the following, we present a simple, while nontrivial lower bound for $str^{avg,M}(\pi)$ and $str^{avg,E}(\pi)$.

Lemma 6. *For each pair $(\alpha,\beta) \in A$, we have*

$$\Delta(\alpha,\beta) \le d(\sqrt[d]{n}-1), \Delta_E(\alpha,\beta) \le \sqrt{d}(\sqrt[d]{n}-1)$$

Proof: Obviously we can get that both of $\Delta(\alpha,\beta)$ and $\Delta_E(\alpha,\beta)$ can achieve the maximum value at the pair (α',β') as follows:

$$\alpha' = (0,\cdots,0), \beta' = (\sqrt[d]{n}-1,\cdots,\sqrt[d]{n}-1)$$

So we have for each pair $(\alpha,\beta) \in A$:

$$\begin{aligned} \Delta(\alpha,\beta) &\le \Delta(\alpha',\beta') = d(\sqrt[d]{n}-1) \\ \Delta_E(\alpha,\beta) &\le \Delta_E(\alpha',\beta') = \sqrt{d}(\sqrt[d]{n}-1) \end{aligned}$$

■

Proposition 3. *For any SFC π whose domain is a d-dimensional universe U, it must be true that:*

$$str^{avg,M}(\pi) \ge \frac{1}{3d}\frac{n+1}{\sqrt[d]{n}-1}$$

and

$$str^{avg,E}(\pi) \ge \frac{1}{3\sqrt{d}}\frac{n+1}{\sqrt[d]{n}-1}$$

Proof: We first prove for $str^{avg,M}(\pi)$. From Lemma 6 we have:

$$\frac{2}{n(n-1)} \sum_{(\alpha,\beta)\in A} \frac{\Delta_\pi(\alpha,\beta)}{\Delta(\alpha,\beta)} \ge \frac{2}{n(n-1)d(\sqrt[d]{n}-1)} \sum_{(\alpha,\beta)\in A} \Delta_\pi(\alpha,\beta)$$

From Lemma 2, we already have:

$$S_{A'}(\pi) = \frac{1}{3}(n-1)n(n+1)$$

Note that A' is the set of all possible ordered pairs in U while A is the set of all possible unordered pairs in U. It follows that:

$$\sum_{(\alpha,\beta)\in A} \Delta_\pi(\alpha,\beta) = \frac{1}{2}S_{A'}(\pi) = \frac{1}{6}(n-1)n(n+1)$$

Combining with the inequality above we have:

$$str^{avg,M}(\pi) = \frac{2}{n(n-1)} \sum_{(\alpha,\beta)\in A} \frac{\Delta_\pi(\alpha,\beta)}{\Delta(\alpha,\beta)} \ge \frac{1}{3d}\frac{n+1}{\sqrt[d]{n}-1}$$

Following the same steps we can prove $str^{avg,E}(\pi) \ge \frac{1}{3\sqrt{d}}\frac{n+1}{\sqrt[d]{n}-1}$ as well. ■

In the following, we show that an upper bound on the performance of the simple curve S defined in 8.

Proposition 4. *If S is the d dimensional simple curve, then*

$$str^{avg,M}(S) \le n^{1-\frac{1}{d}}, str^{avg,E}(S) \le \sqrt{2}n^{1-\frac{1}{d}}$$

Lemma 7. *Suppose S is the d dimensional simple curve. Then we have for each pair $(\alpha,\beta) \in A$*

$$\frac{\Delta_S(\alpha,\beta)}{\Delta(\alpha,\beta)} \le n^{1-\frac{1}{d}}, \frac{\Delta_S(\alpha,\beta)}{\Delta_E(\alpha,\beta)} \le \sqrt{2}n^{1-\frac{1}{d}}$$

Proof: Arbitrarily choose a pair of cells, say $\alpha = (x_1,\cdots,x_d), \beta = (y_1,\cdots,y_d)$. Then we have:

$$\Delta_S(\alpha,\beta) = |\sum_{i=1}^{d}(x_i-y_i)(\sqrt[d]{n})^{i-1}|, \Delta(\alpha,\beta) = \sum_{i=1}^{d}|x_i-y_i|$$

It follows that:

$$\begin{aligned} \frac{\Delta_S(\alpha,\beta)}{\Delta(\alpha,\beta)} &\le \frac{\sum_{i=1}^{d}|(x_i-y_i)|(\sqrt[d]{n})^{i-1}}{\sum_{i=1}^{d}|x_i-y_i|} \\ &\le \frac{(\sqrt[d]{n})^{d-1}\sum_{i=1}^{d}|(x_i-y_i)|}{\sum_{i=1}^{d}|x_i-y_i|} \\ &= (\sqrt[d]{n})^{d-1} \end{aligned}$$

Note that:

$$\Delta_E(\alpha,\beta) = \sqrt{\sum_{i=1}^{d}|x_i-y_i|^2} \ge \frac{1}{\sqrt{2}}\sum_{i=1}^{d}|x_i-y_i| = \frac{1}{\sqrt{2}}\Delta(\alpha,\beta)$$

It follows that:

$$\begin{aligned} \frac{\Delta_S(\alpha,\beta)}{\Delta_E(\alpha,\beta)} &\le \sqrt{2}\frac{\Delta_S(\alpha,\beta)}{\Delta(\alpha,\beta)} \\ &\le \sqrt{2}(\sqrt[d]{n})^{d-1} \end{aligned}$$

■

Now we are ready to prove Proposition 4.

Proof: Since $str^{avg,M}(S)$ and $str^{avg,E}(S)$ is the average all pairs stretch by using Manhattan metric and Euclidean metric respectively, we can get the inequalities in Proposition 4 directly from the results in Lemma 7. ■

VI. Conclusions

We presented an analysis of the proximity preserving properties of an SFC. We showed that there is an inherent lower bound on the average-average NN-stretch of any SFC. Further, some specific SFCs such as the Z curve come close to matching this bound. This study raises a number of interesting open questions.

- An obvious question is to close the gap between the lower bound and upper bound for the average-average NN-stretch, perhaps via an analysis of a different SFC, or through a better lower bound, or both. A related question is an analysis of the average NN-stretch of the Hilbert SFC.

- Next, there is a larger gap between the lower bound and the upper bound for the average-maximum NN-stretch. It would be interesting to narrow this gap.
- It would also be interesting to narrow the gap between upper and lower bounds for the all pairs stretch.
- Another direction is the analysis of proximity preservation using a more general probabilistic model of input.

References

[1] D. Abel and D. Mark. A comparative analysis of some two-dimensional orderings. *Int. J. Geogr. Inf. Syst.*, 4(1):21–31, 1990.

[2] J. Alber and R. Niedermeier. On multi-dimensional hilbert indexings. In *Proc. 4th International Conference on Computing and Combinatorics (COCOON)*, pages 329–338, 1998.

[3] S. Aluru and F. Sevilgen. Parallel domain decomposition and load balancing using space-filling curves. In *High-Performance Computing, 1997. Proceedings. Fourth International Conference on*, pages 230 –235, 1997.

[4] W. G. Aref and I. Kamel. On multi-dimensional sorting orders. In *Database and Expert Systems Applications (DEXA)*, pages 774–783, 2000.

[5] H.-L. Chen and Y.-I. Chang. Neighbor-finding based on space-filling curves. *Inf. Syst.*, 30:205–226, May 2005.

[6] A. Collins, J. Czyzowicz, L. Gasieniec, and A. Labourel. Tell me where I am so I can meet you sooner. In *Proc. ICALP*, pages 502–514, 2010.

[7] H.-K. Dai and H.-C. Su. On the locality properties of space-filling curves. In *Proc. International Symposium on Algorithms and Computation (ISAAC)*, pages 385–394, 2003.

[8] H.-K. Dai and H.-C. Su. On p-norm based locality measures of space-filling curves. In *Proc. International Symposium on Algorithms and Computation (ISAAC)*, pages 364–376, 2004.

[9] C. Faloutsos. Multiattribute hashing using gray codes. *SIGMOD Record*, 15:227–238, June 1986.

[10] C. Faloutsos. Gray codes for partial match and range queries. *IEEE Trans. Software Engg.*, 14:1381–1393, 1988.

[11] C. Gotsman and M. Lindenbaum. On the metric properties of discrete space-filling curves. *IEEE Transactions on Image Processing*, 5(5):794 –797, 1996.

[12] L. H. Harper. Optimal assignments of numbers to vertices. *Journal of the Society for Industrial and Applied Mathematics*, 12(1):131–135, 1964.

[13] D. Hilbert. Uber die stegie abbildung einer linie auf flachenstuck. *Math. Ann.*, 38:459–460, 1891.

[14] H. V. Jagadish. Analysis of the hilbert curve for representing two-dimensional space. *Information Processing Letters*, 62:17–22, 1997.

[15] G. Jin and J. M. Mellor-Crummey. Sfcgen: A framework for efficient generation of multi-dimensional space-filling curves by recursion. *ACM Transactions Mathematical Software*, 31(1):120–148, 2005.

[16] Y. Matias and A. Shamir. A video scrambling technique based on space filling curves. In *CRYPTO*, pages 398–417, 1987.

[17] G. Mitchison and R. Durbin. Optimal numberings of an n x n array. *SIAM J. Algebraic Discrete Methods*, 7:571–582, October 1986.

[18] B. Moon, H. V. Jagadish, C. Faloutsos, and J. H. Saltz. Analysis of the clustering properties of the hilbert space-filling curve. *IEEE Trans. Knowledge and Data Engineering*, 13(1):124–141, 2001.

[19] G. Morton. A computer oriented geodetic data base; and a new technique in file sequencing. Technical report, IBM, 1966.

[20] R. Niedermeier, K. Reinhardt, and P. Sanders. Towards optimal locality in mesh-indexings. *Discrete Applied Mathematics*, 117(1-3):211–237, 2002.

[21] J. A. Orenstein and T. H. Merrett. A class of data structures for associative searching. In *Proceedings of the 3rd ACM SIGACT-SIGMOD symposium on Principles of database systems*, PODS '84, pages 181–190, 1984.

[22] M. Parashar and J. C. Browne. On partitioning dynamic adaptive grid hierarchies. In *Proceedings of the 29th Annual Hawaii International Conference on System Sciences*, pages 604–613, 1996.

[23] J. R. Pilkington and S. B. Baden. Dynamic partitioning of non-uniform structured workloads with space filling curves. *IEEE Trans. on Parallel and Distributed Systems*, 7(3):288 – 300, 1996.

[24] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, 1990.

[25] S. Tirthapura, S. Seal, and S. Aluru. A formal analysis of space filling curves for parallel domain decomposition. In *Proc. International Conference on Parallel Processing (ICPP)*, pages 505–512, 2006.

[26] M. Warren and J. Salmon. A parallel hashed-octtree N-body algorithm. In *Proceedings of Supercomputing '93*, Portland, OR, Nov. 1993.