On Optimality of Clustering by Space Filling Curves

Pan Xu
panxu@iastate.edu

Srikanta Tirthapura
snt@iastate.edu

Iowa State University

Multi-dimensional Data

- Indexing and managing single-dimensional data is a solved problem

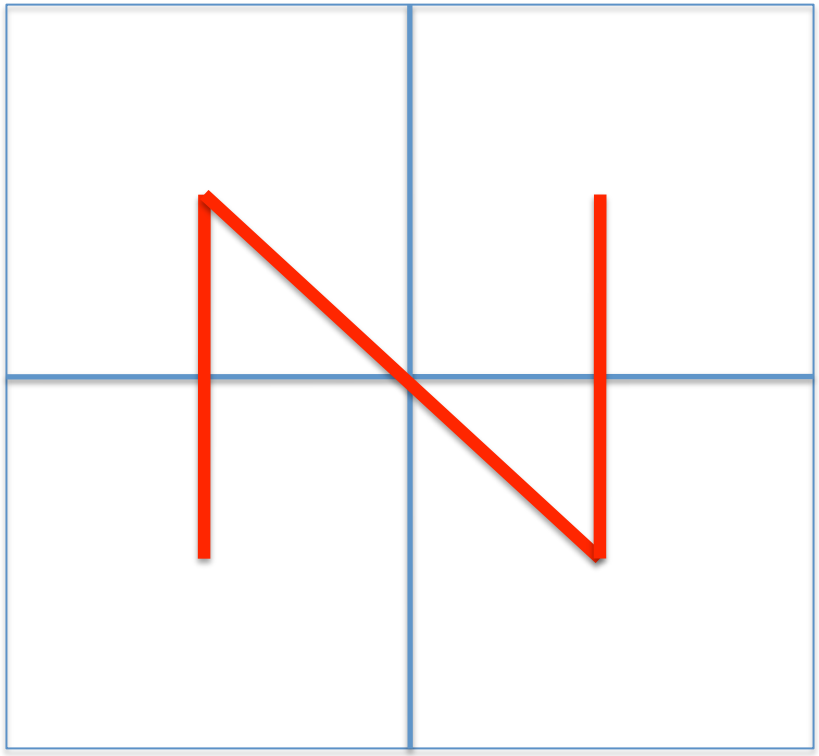
- Multi-dimensional data is not
  - Computational Geometry (k-d trees, quad trees)
  - Databases (R-trees, Space Filling Curves,etc)
  - Parallel Computing (same as above)

Space Filling Curve (SFC)

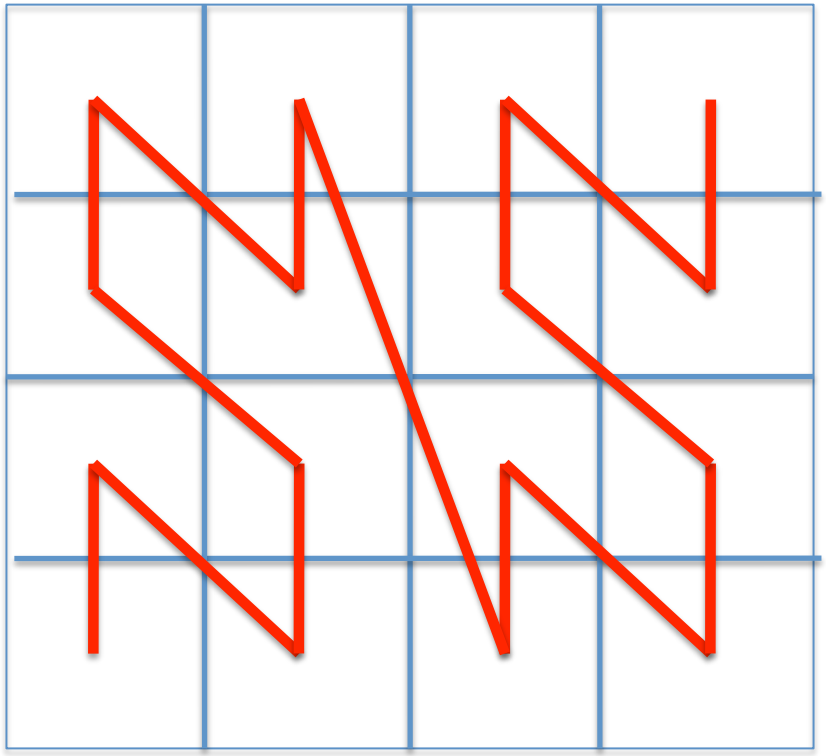
- Discrete multi-dimensional universe
- Path that passes through each point in a multi-dimensional universe exactly once
- Recipe for processing multi-dimensional data
  - Use and SFC to map data to a single dimension
  - Use a single dimensional data structure for indexing (ex: B+ trees)

Z Space Filling Curve

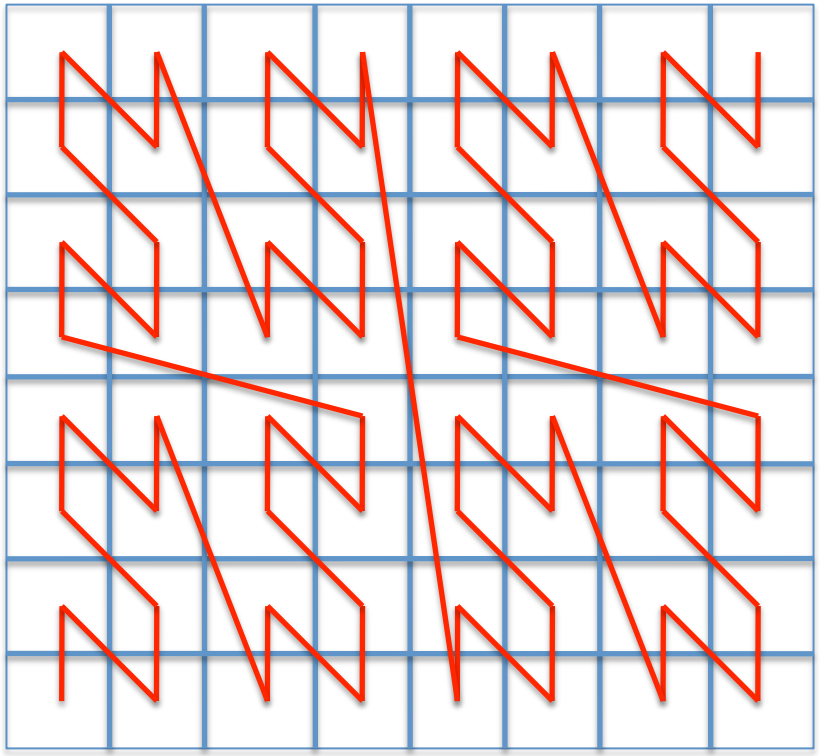
Z Space Filling Curve



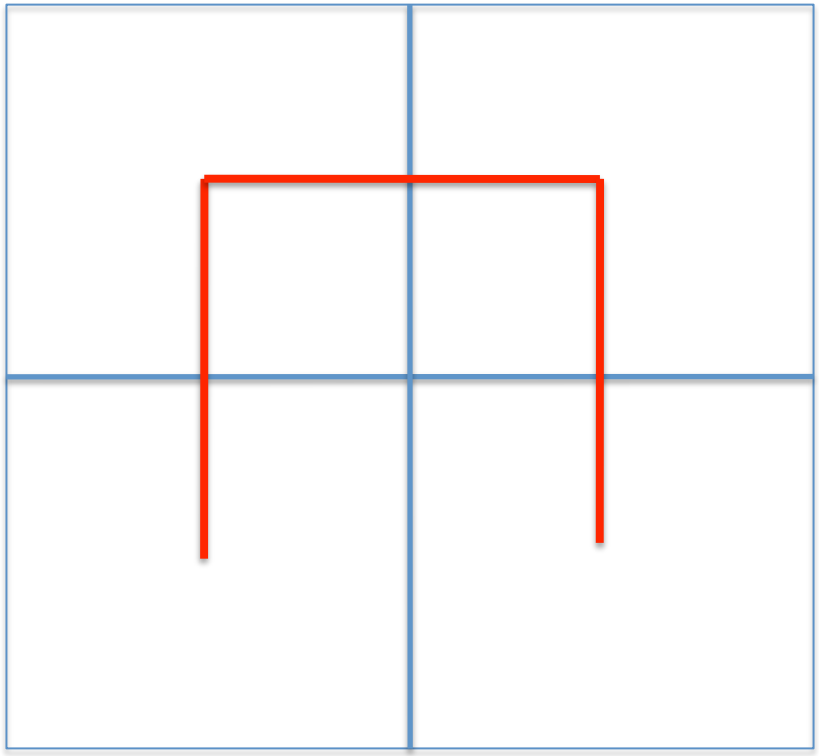
Z Space Filling Curve



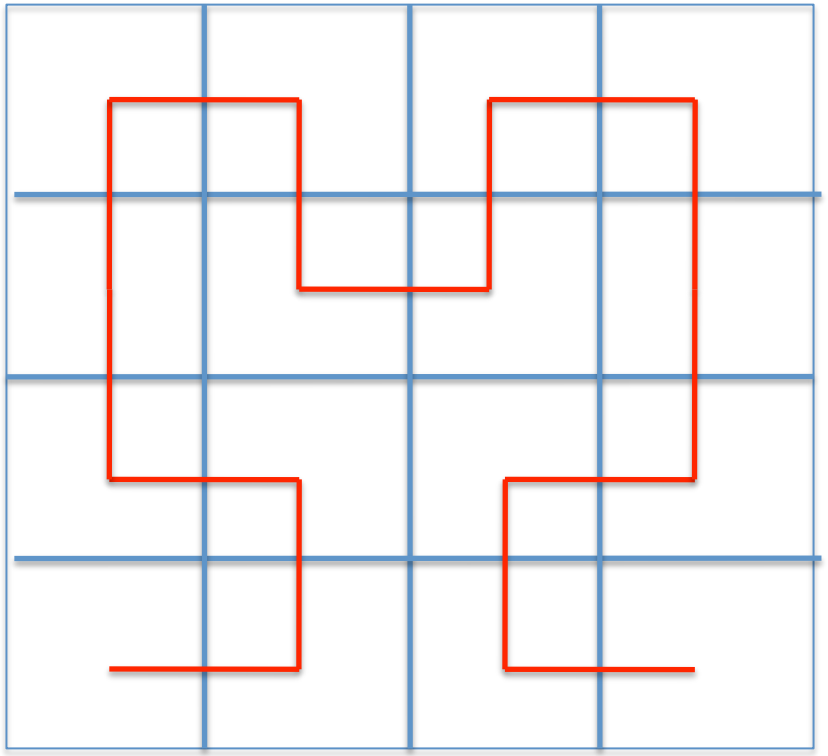
Z Space Filling Curve



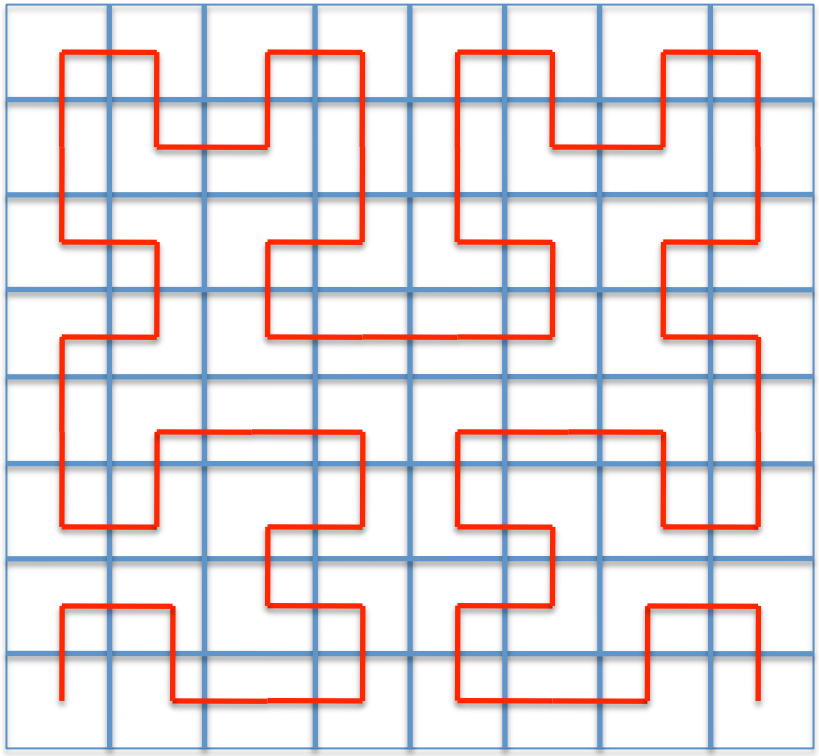
Hilbert Space Filling Curve



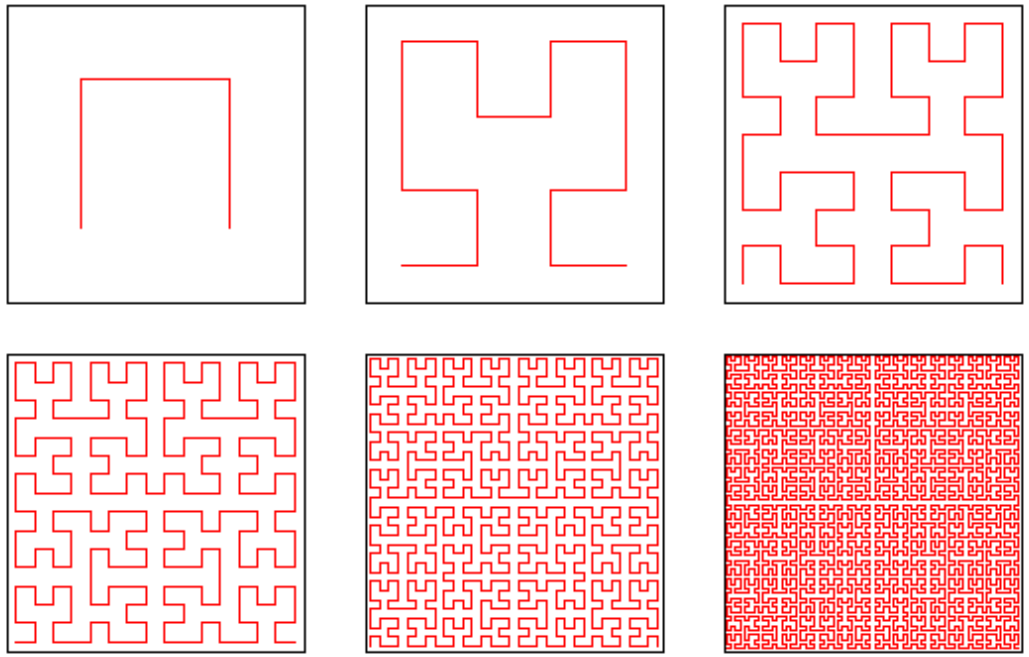
Hilbert Space Filling Curve



Hilbert Space Filling Curve



Hilbert Space Filling Curve



Source: http://www.math.osu.edu/~fiedorowicz.1/math655/Peano.html
Prof. Zbigniew Fiedorowicz

Where SFCs?

- Databases
  - Spatial databases, Geographic Information Systems
  - Commercial and Open Source Products: Oracle Spatial, OpenGIS
- Parallel Computing
  - Partitioning data among processors
  - Deciding which data to include inside a cache
- Thousands of Citations

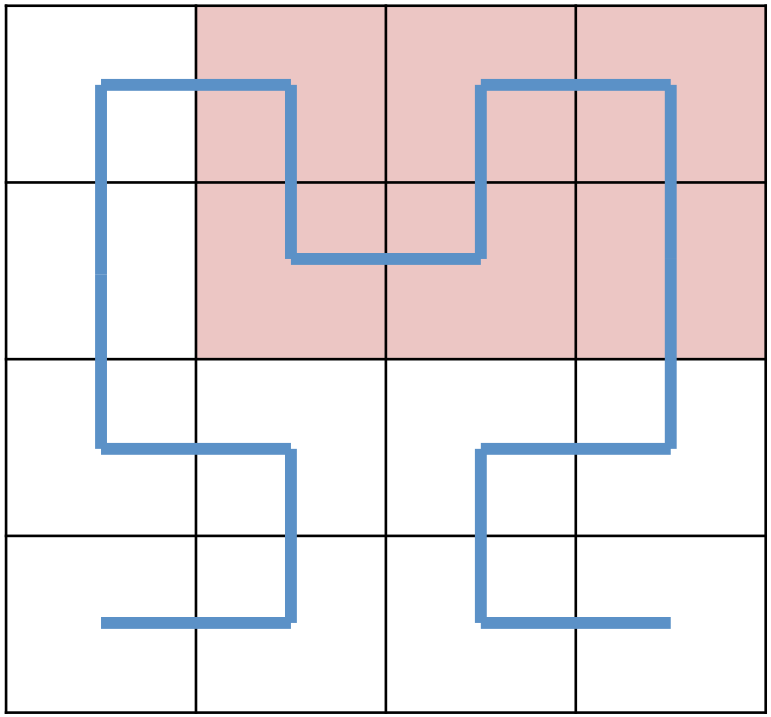
Performance of an SFC

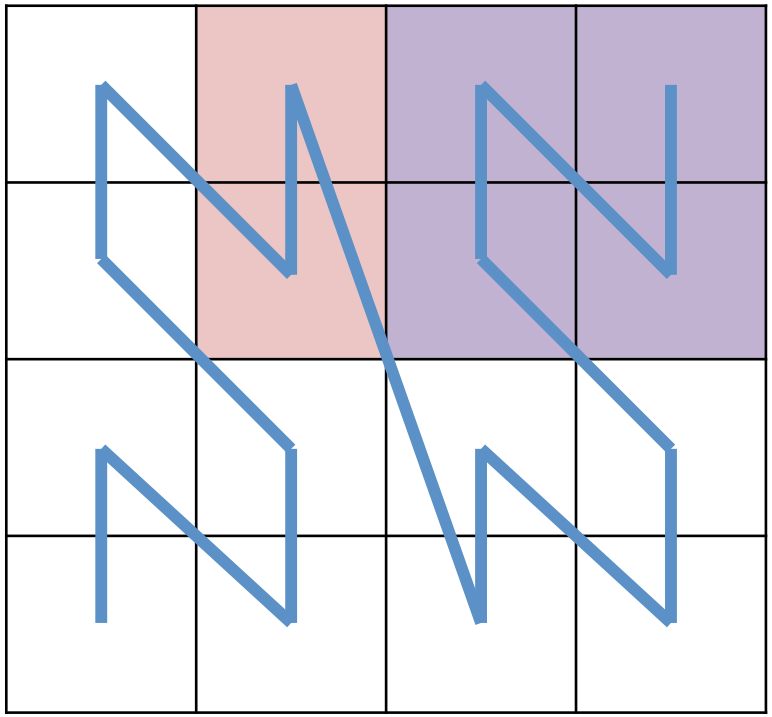
- Range Queries
  - Find me all coffee shops between latitudes x and x', longitudes y and y', and priced between z and z'
  - A query is a subset of the universe
    - Rectangular Queries: Intersection of halfplanes
    - Rectilinear Queries: Union of multiple disjoint rectangles
- Using SFC, a multidimensional range split into many one dimensional ranges
- Each one dimensional range is evaluated "quickly" using an index

Clustering Number

- Given a query q and SFC π, the clustering number π of SFC for query q is defined as **the smallest number of contiguous regions on the SFC that the query q can be divided into**

Clustering Number





1 for the Hilbert Curve

2 for the Z Curve

Average Clustering Number

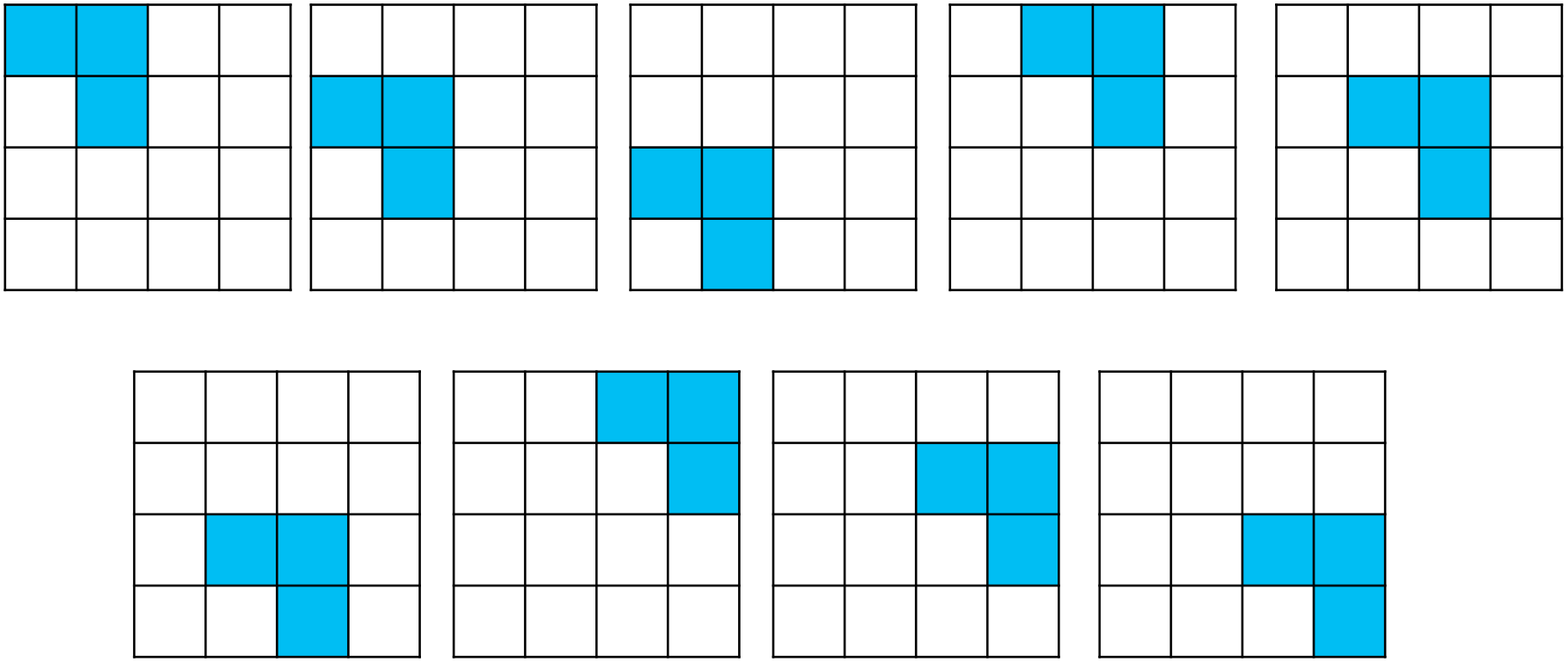
- For a set of queries Q, the average clustering number of an SFC π is:

$$c(Q,\pi) = \frac{\sum_{q \in Q} c(q,\pi)}{|Q|}$$

- Set of queries formed by:
  - Translations of a basic query shape
  - Rotations of a query
  - Translations + Rotations

Translation of a Query Shape

Q = T(q) = all translations of a basic query shape q



Rotation

With d dimensions, d! possible rotations for a single query.

Questions

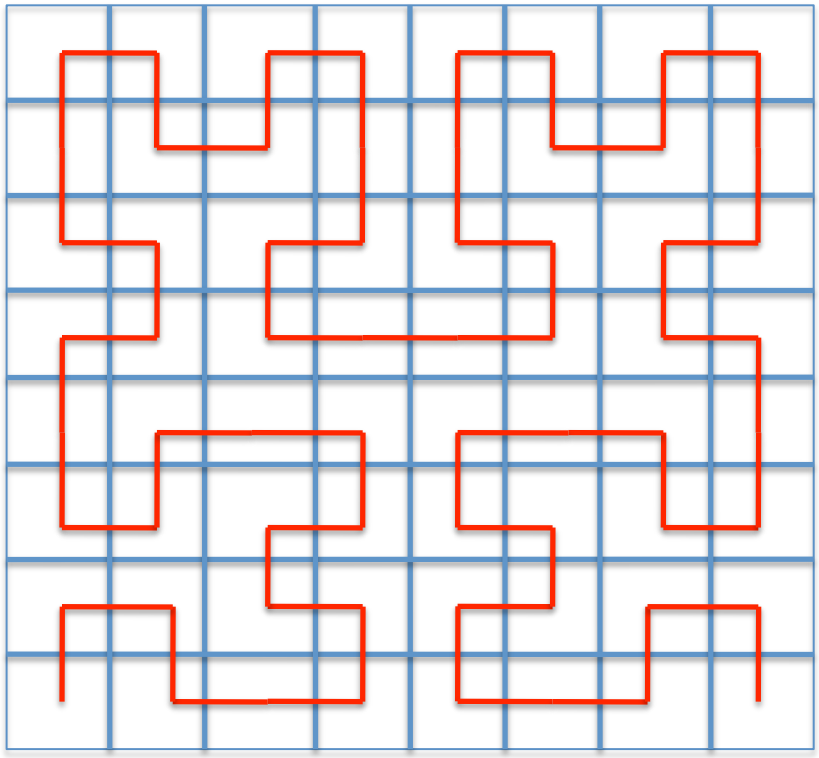
- Given SFC π and query set Q, what is c(Q,π)?
- Lower bounds on c(Q,π)?
- Optimality of a curve for a class of queries?

- Setting:
  - d dimensional universe $n^{1/d} \times n^{1/d} \times \ldots$
  - Total number of cells = n

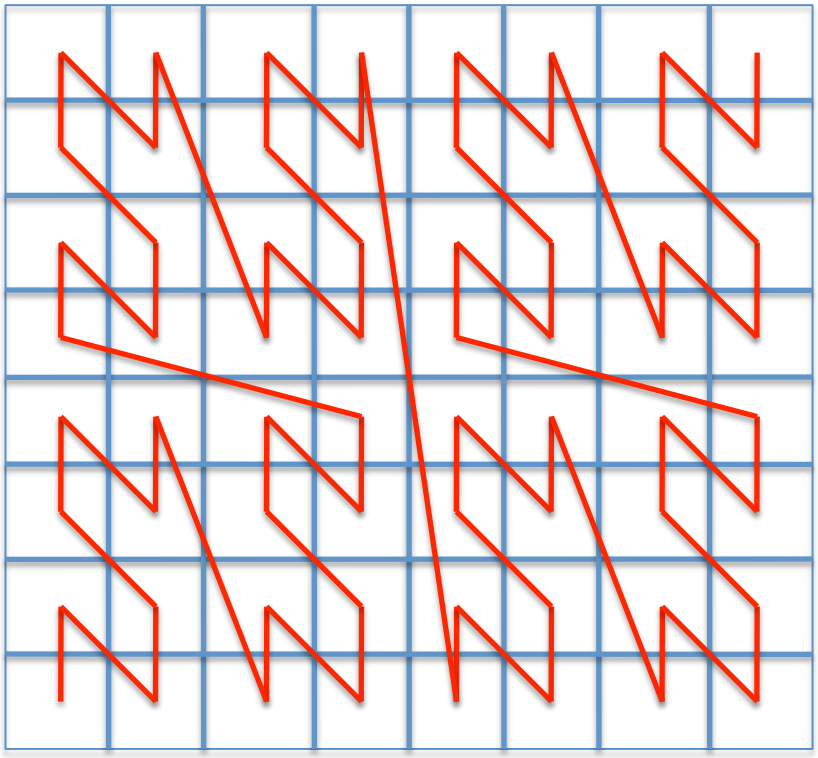
Prior Work

- Orenstein and Merrett (1984)
  - proposed use of SFCs for use in range queries for databases
- Jagadish (1997)
  - In two dimensions, the clustering number of the Hilbert curve for 2 x 2 queries is 2
- Moon, Jagadish, Faloutsos, Saltz (2001)
  - In d dimensions, the clustering number of the Hilbert curve on rectilinear queries is $\frac{S_g}{2d}$ where s_g is the surface area of query *g*, and *d* is the number of dimensions
- 300+ citations to the above work since 2001

Continuous SFCs



Continuous SFC



Non-Continuous SFC

Our Results

	Continuous SFCs	General SFCs
Rectangular Queries	Exact Formula	Lower Bound, Continuous is optimal
General (Rectilinear) Queries	Exact Formula	Continuous need not be optimal

Our Results

- When all rotations are considered for rectangular queries, Hilbert SFC is optimal (and so is any continuous SFC)

- Note:
  - Results for limiting *n*, and for constant query size, which does not increase with *n*

General Techniques

For points α, β and query q, define 0-1 function $I(q, \alpha, \beta)$

- 1 if $\alpha \in q, \beta \in q$
- 0 otherwise

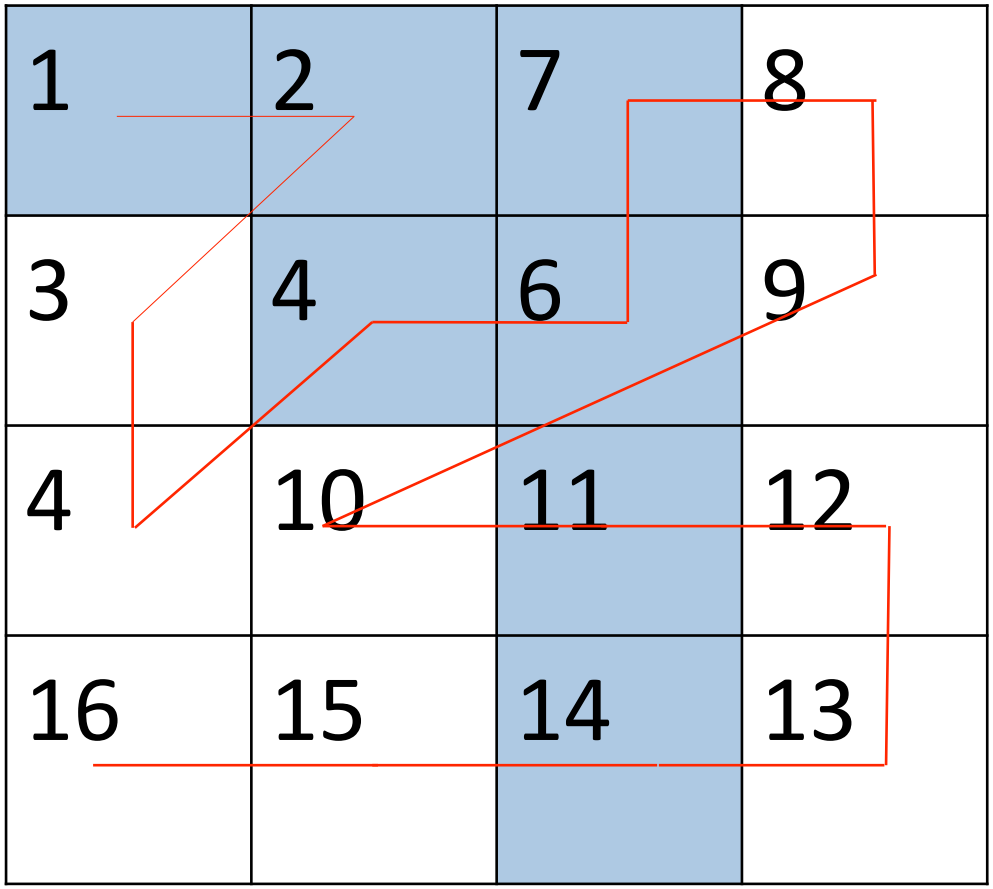
Basic Lemma

- $N(\pi)$ = All pairs of vertices in SFC π

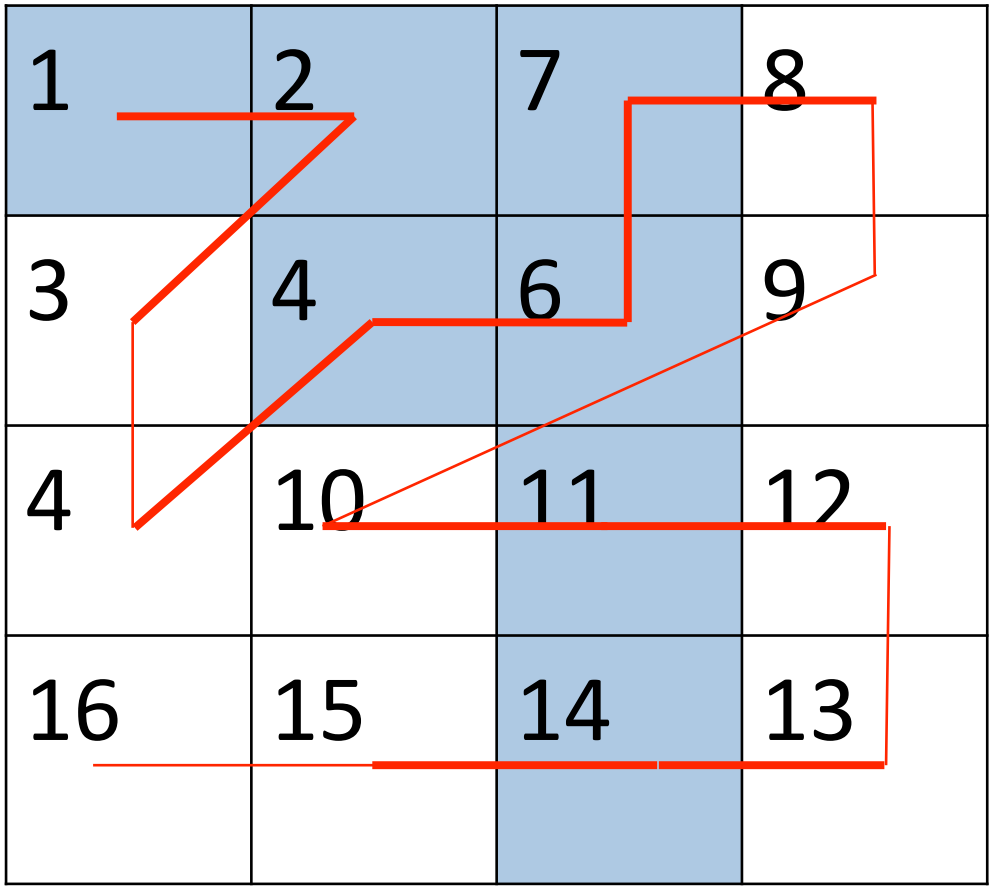
- Lemma: For any query q,

$$c(q,\pi) = |q| - \sum_{(\alpha,\beta)\in N(\pi)} I(q,\alpha,\beta)$$

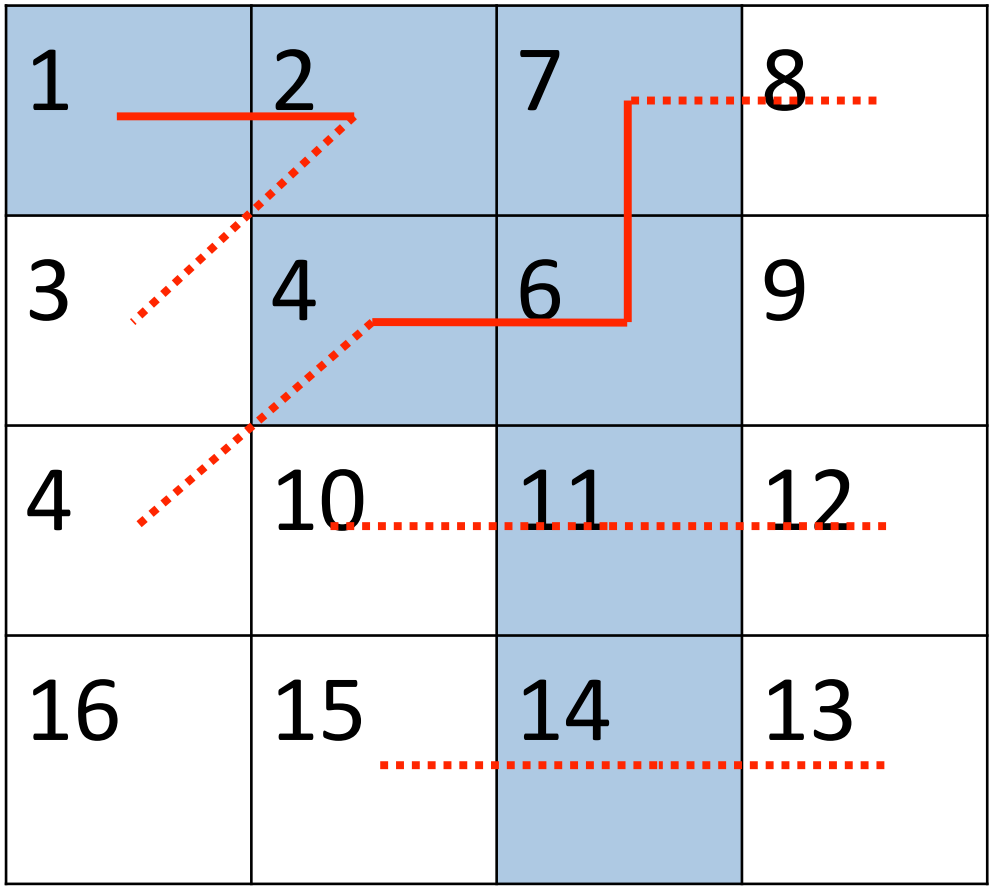
Basic Lemma



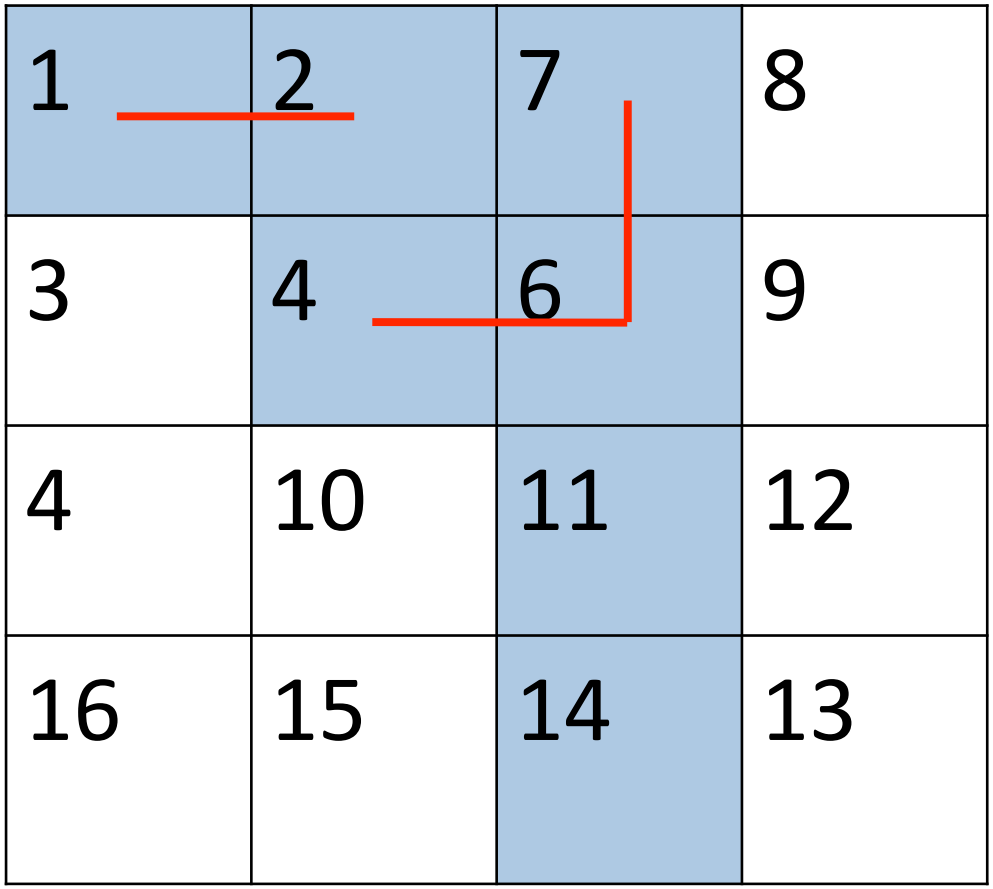
Basic Lemma



Basic Lemma



Basic Lemma



Average Clustering Number

- For each edge (α, β) that is part of the SFC, and set of queries Q, let

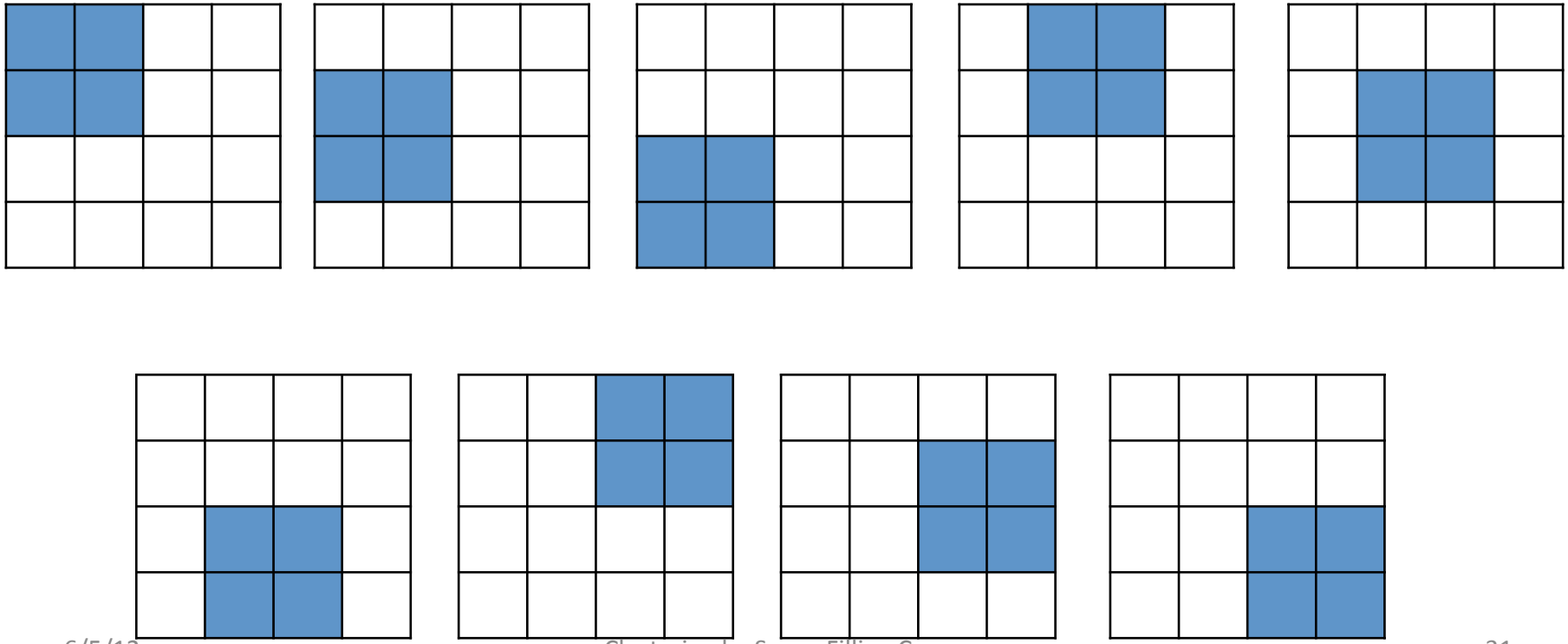
$$P_Q(\alpha, \beta) = \{\, q \in Q \mid I(q, \alpha, \beta) = 1\}$$

- Lemma: Suppose Q is the set of queries formed by transformations of a basic shape q

$$c(Q, \pi) = |q| - \frac{\sum_{(\alpha,\beta)\in\pi} P_Q(\alpha, \beta)}{|Q|}$$

Counting Question

- For a set of queries Q and cells α, β, how many queries in Q include both α and β?



Continuous SFC, Arbitrary Query, Translations Only

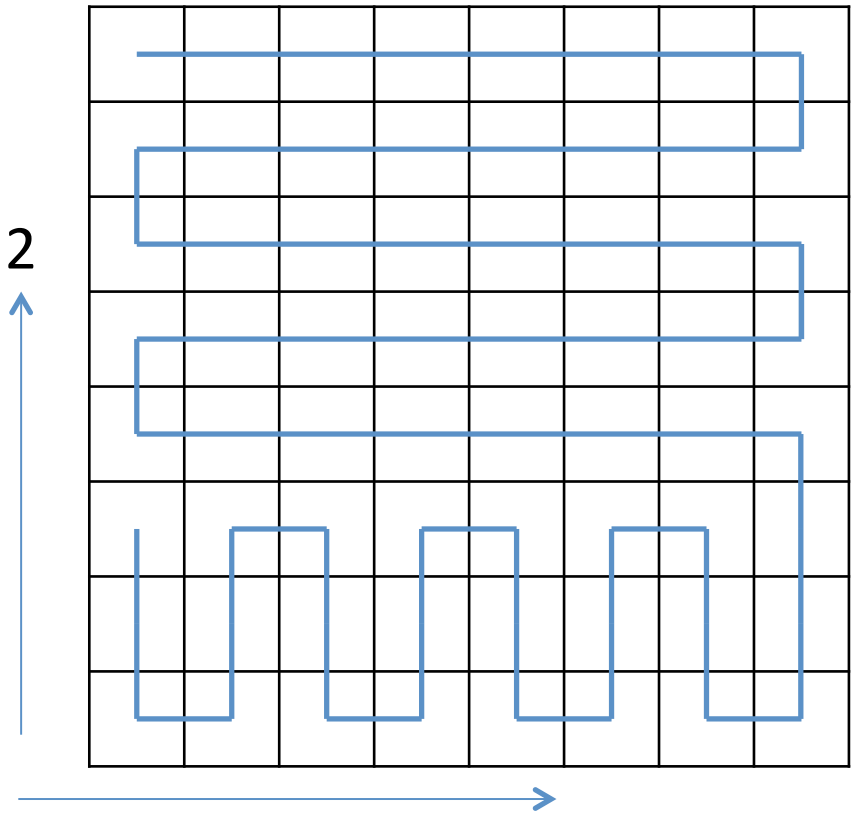
- Given query shape g, and SFC π

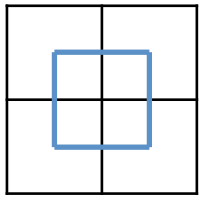
$$\lim_{n\to\infty} c(T(g),\pi) = |g| - \mu(\pi).\nu(g)$$

- $\mu(\pi) = [\mu_1(\pi), \mu_2(\pi), \ldots.]$
  $\mu_i(\pi)$ is fraction of edges of π along dim. i

- $\nu(g) = [\nu_1(g), \nu_2(g), \ldots]$
  $\nu_i(g)$ number of edges in "skeleton" of g along dim. i

Continuous SFC

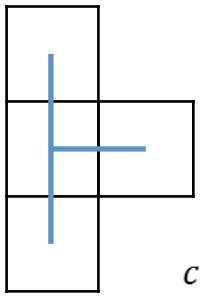




$$v(A) = [2,2]$$

$$c(T(A), \pi) = 4 - 2 = 2$$

Query A



$$v(B) = [1,2]$$

$$c(T(B), \pi) = 4 - \left(\frac{5}{8} \cdot 1\right) - \left(\frac{3}{8} \cdot 2\right)$$

Query B

$$\mu(\pi) = \left[\frac{5}{8}, \frac{3}{8}\right]$$

Continuous SFC, Arbitrary Query

- Above Formula allows exact computation of average clustering number

- Also extends to Translation + Rotation

Symmetric SFC, Arbitrary Query

- A symmetric SFC is one which has the same number of edges along each dimension i, i.e.

$$\mu(\pi) = \left[\frac{1}{d}, \frac{1}{d}, \dots\right]$$

- Corollary: For every symmetric SFC π, and query shape g,

$$c(T(g), \pi) = \frac{S_g}{2d}$$

where S_g is the surface area of g

- The result of Moon et al. (2001) follows from the above

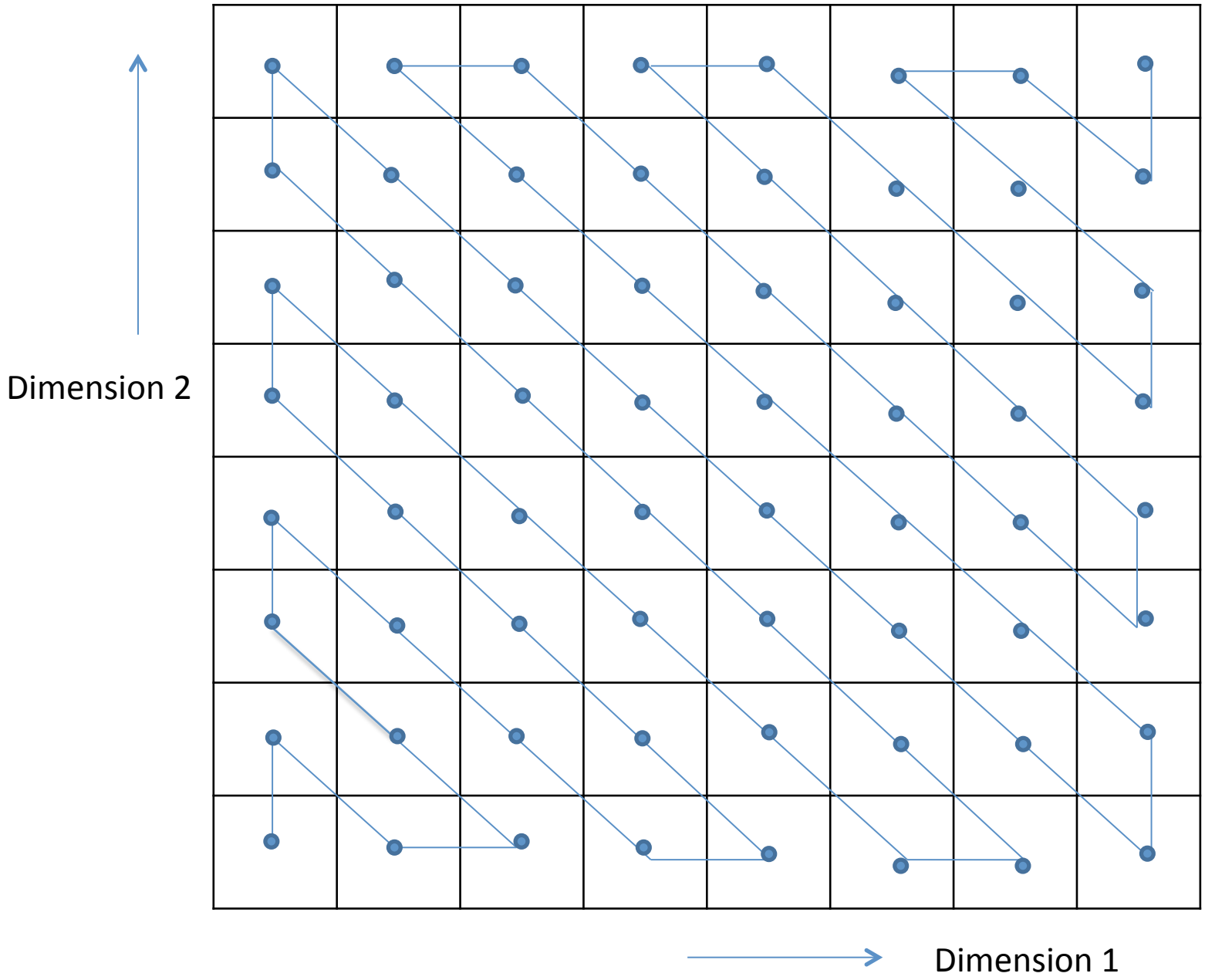
Arbitrary SFC, Rectangular Query

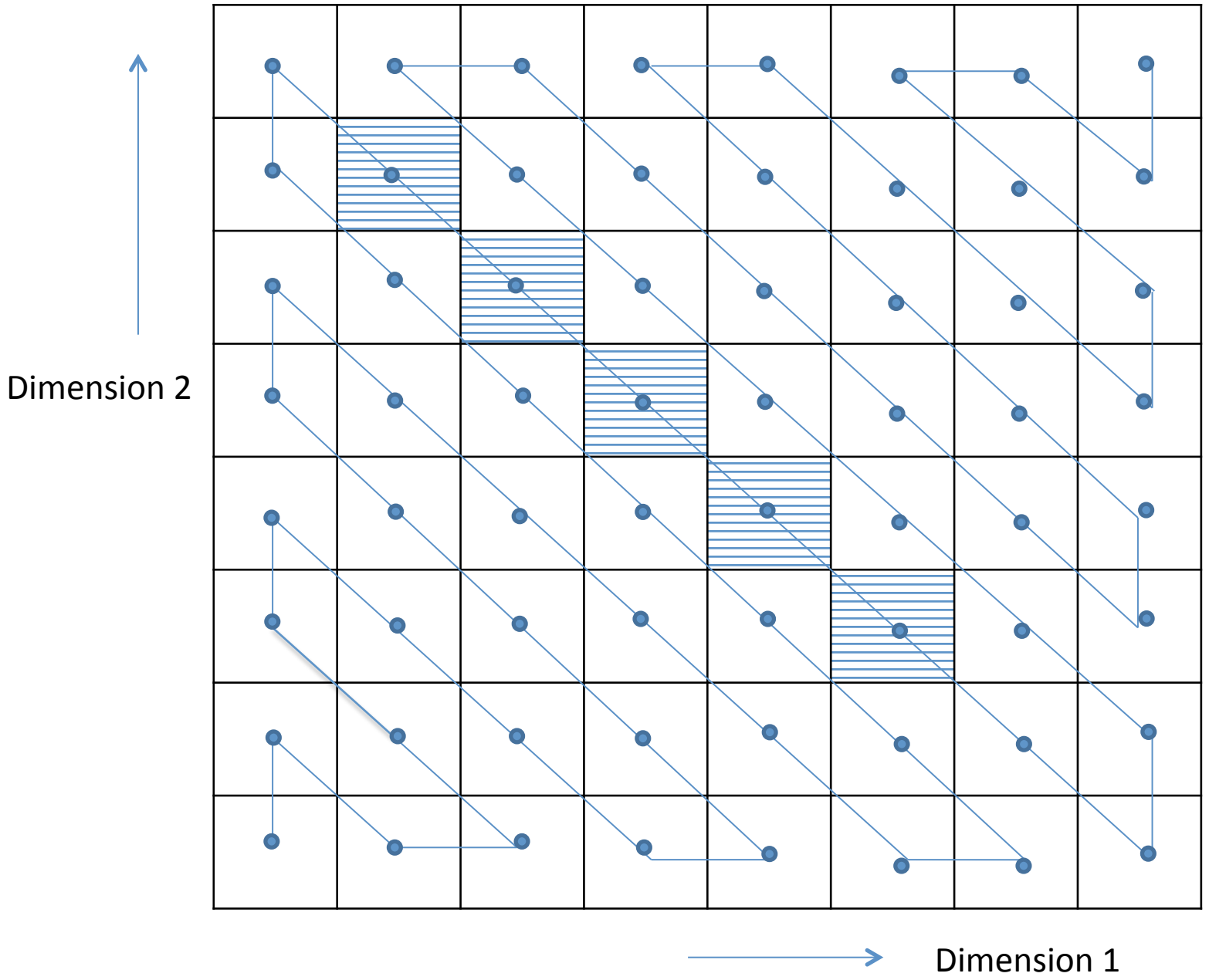
- Lower Bound Theorem: For any rectangular query r, and any subset of rotations $\Lambda \subseteq \Lambda^*$, for any SFC π (not necessarily continuous)

$$\lim_{n\to\infty} c(Q(r,\Lambda),\pi) \geq |r| - \nu^{max}$$

Where ν^{max} is the maximum element of corresponding ν vector for rotations.

- Corollary: For any rectangular query r, and subset of rotations, there is a continuous SFC that is optimal


Dimension 2
Dimension 1


Dimension 2
Dimension 1

Our Results

	Continuous SFCs	General SFCs
Rectangular Queries	Exact Formula	Lower Bound, Continuous is optimal
General (Rectilinear) Queries	Exact Formula	Continuous need not be optimal

Conclusions

- Fundamental Results about SFCs
- General Method from first principles
- “Edge-centric” perspective, i.e. count the number of queries crossing the edges
- Answers Explicit Open Questions by Jagadish (1997), Moon et al. (2001) and others