



## 15-826: Multimedia Databases and Data Mining

Lecture #11: Fractals: M-trees and dim.  
curse (case studies – Part II)

C. Faloutsos

---

---

---

---

---

---



## Must-read Material

- Alberto Belussi and Christos Faloutsos,  
[Estimating the Selectivity of Spatial Queries  
Using the 'Correlation' Fractal Dimension](#)  
Proc. of VLDB, p. 299-310, 1995

15-826

Copyright: C. Faloutsos (2011)

2

---

---

---

---

---

---



## Optional Material

Optional, but **very** useful: Manfred Schroeder  
*Fractals, Chaos, Power Laws: Minutes  
from an Infinite Paradise* W.H. Freeman  
and Company, 1991

15-826

Copyright: C. Faloutsos (2011)

3

---

---

---

---

---

---



## Outline

Goal: 'Find similar / interesting things'

- Intro to DB
- ➡ • Indexing - similarity search
- Data Mining

15-826

Copyright: C. Faloutsos (2011)

4

---

---

---

---

---

---

---



## Indexing - Detailed outline

- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
  - z-ordering
  - R-trees
  - misc
- ➡ • fractals
  - intro
  - applications
- text

15-826

Copyright: C. Faloutsos (2011)

5

---

---

---

---

---

---

---



## Indexing - Detailed outline

- fractals
  - intro
  - applications
    - disk accesses for R-trees (range queries)
    - dimensionality reduction
  - ➡ • selectivity in M-trees
    - dim. curse revisited
    - “fat fractals”
    - quad-tree analysis [Gaedé+]

15-826

Copyright: C. Faloutsos (2011)

6

---

---

---

---

---

---

---



## What else can they solve?

- ✓ separability [KDD'02]
  - forecasting [CIKM'02]
- ✓ dimensionality reduction [SBB'D'00]
  - non-linear axis scaling [KDD'02]
- ✓ disk trace modeling [Wang+'02]
- ➡ selectivity of spatial/multimedia queries  
[PODS'94, VLDB'95, ICDE'00]
- ...

15-826

Copyright: C. Faloutsos (2011)

7

---



---



---



---



---



---



---



---



## Metric trees - analysis

- Problem: How many disk accesses, for an M-tree?
- Given:
  - N (# of objects)
  - C (fanout of disk pages)
  - r (radius of range query - BIASED model)

15-826

Copyright: C. Faloutsos (2011)

8

---



---



---



---



---



---



---



---



## Metric trees - analysis

- Problem: How many disk accesses, for an M-tree?
- Given:
  - N (# of objects)
  - C (fanout of disk pages)
  - r (radius of range query - BIASED model)
- NOT ENOUGH - what else do we need?

15-826

Copyright: C. Faloutsos (2011)

9

---



---



---



---



---



---



---



---



## Metric trees - analysis

- A: something about the distribution

15-826

Copyright: C. Faloutsos (2011)

10

---



---



---



---



---



---



---



---



---



---



## Metric trees - analysis

- A: something about the distribution
- [Ciaccia, Patella, Zezula, PODS98]: assumed that the distance distribution is the same, for every object:



Paolo Ciaccia



Marco Patella

15-826

Copyright: C. Faloutsos (2011)

11

---



---



---



---



---



---



---



---



---



---



## Metric trees - analysis

- A: something about the distribution
- [Ciaccia+, PODS98]: assumed that the distance distribution is the same, for every object:
- $$F_1(d) = \text{Prob}(\text{an object is within } d \text{ from object \#1})$$
- $$= F_2(d) = \dots = F(d)$$

15-826

Copyright: C. Faloutsos (2011)

12

---



---



---



---



---



---



---



---



---



---



## Metric trees - analysis

- A: something about the distribution
- Given our ‘fractal’ tools, we could try them - which one?

15-826

Copyright: C. Faloutsos (2011)

13

---



---



---



---



---



---



---



---



---



## Metric trees - analysis

- A: something about the distribution
- Given our ‘fractal’ tools, we could try them - which one?
- A: Correlation integral [Traina+, ICDE2000]

15-826

Copyright: C. Faloutsos (2011)

14

---



---



---



---



---



---



---



---

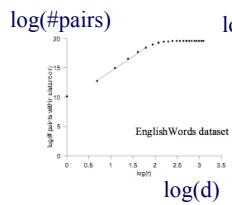


---

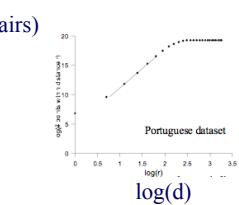


## Metric trees - analysis

English dictionary



Portuguese dictionary



15-826

Copyright: C. Faloutsos (2011)

15

---



---



---



---



---



---



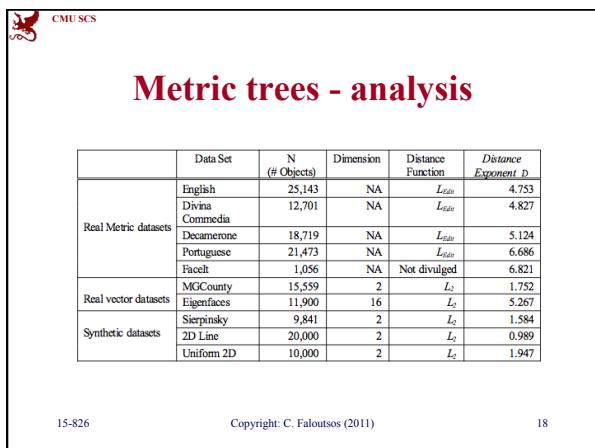
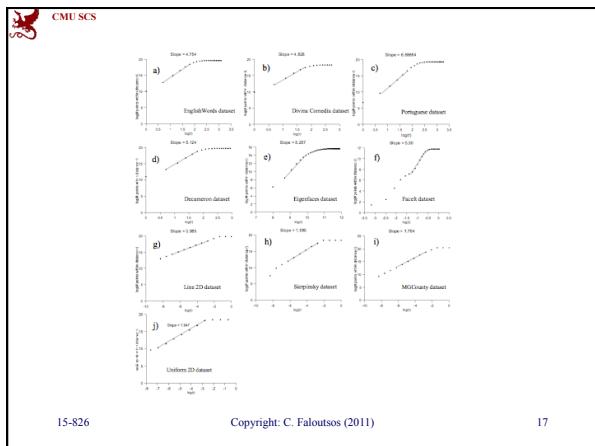
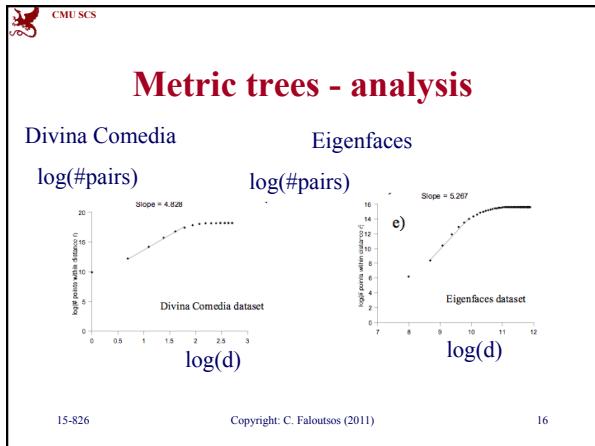
---



---



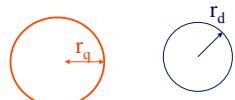
---





## Metric trees - analysis

- So, what is the # of disk accesses, for a node of radius  $r_d$ , on a query of radius  $r_q$ ?



15-826

Copyright: C. Faloutsos (2011)

19

---

---

---

---

---

---

---

---

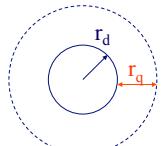
---

---



## Metric trees - analysis

- So, what is the # of disk accesses, for a node of radius  $r_d$ , on a query of radius  $r_q$ ?
- A:  $\sim (r_d + r_q) \dots$



15-826

Copyright: C. Faloutsos (2011)

20

---

---

---

---

---

---

---

---

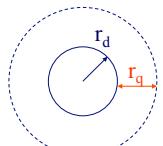
---

---



## Metric trees - analysis

- So, what is the # of disk accesses, for a node of radius  $r_d$ , on a query of radius  $r_q$ ?
- A:  $\sim (r_d + r_q)^D$



15-826

Copyright: C. Faloutsos (2011)

21

---

---

---

---

---

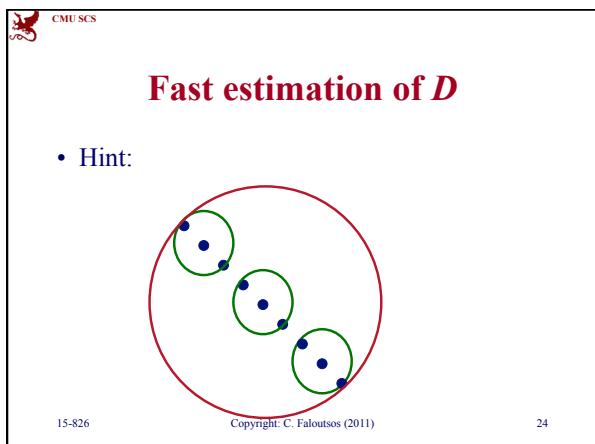
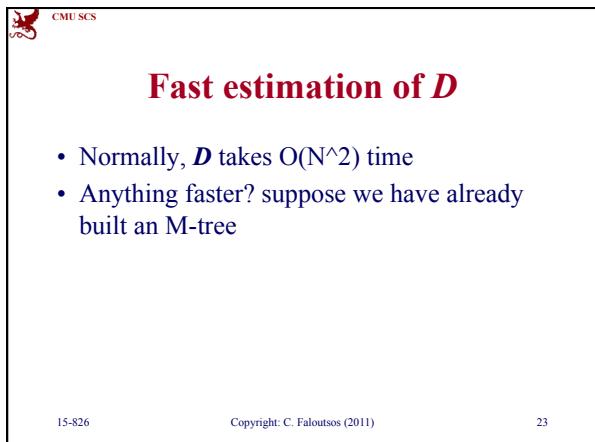
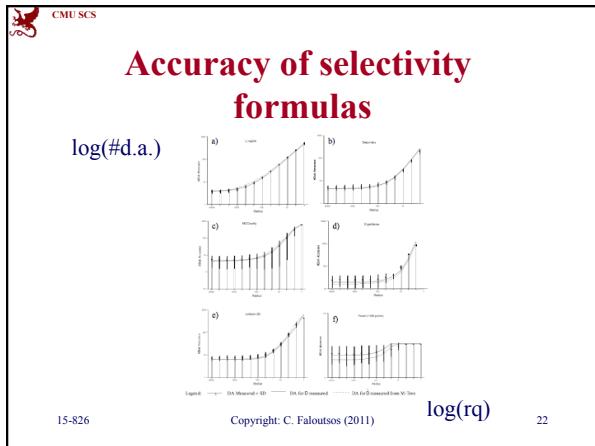
---

---

---

---

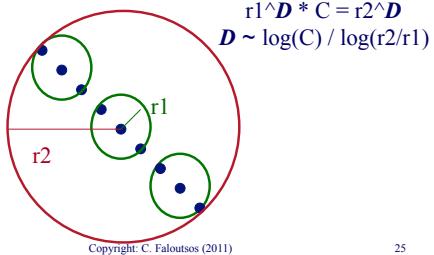
---





## Fast estimation of $D$

- Hint:



15-826

Copyright: C. Faloutsos (2011)

25

---

---

---

---

---

---



## Indexing - Detailed outline

- fractals
  - intro
  - applications
    - disk accesses for R-trees (range queries)
    - dimensionality reduction
    - selectivity in M-trees
    - ➡ • dim. curse revisited
      - “fat fractals”
      - quad-tree analysis [Gaede+]

15-826

Copyright: C. Faloutsos (2011)

26

---

---

---

---

---

---



## Dim. curse revisited

- (Q: how serious is the dim. curse, e.g.:)
- Q: what is the search effort for k-nn?
  - given N points, in E dimensions, in an R-tree, with k-nn queries ('biased' model)

[Pagel, Korn + ICDE 2000]



15-826

Copyright: C. Faloutsos (2011)

27

---

---

---

---

---

---



## (Overview of proofs)

- assume that your points are uniformly distributed in a  $d$ -dimensional manifold (= hyper-plane)
- derive the formulas
- substitute  $d$  for the fractal dimension

15-826

Copyright: C. Faloutsos (2011)

28

---



---



---



---



---



---



---



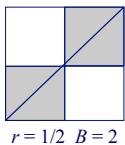
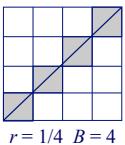
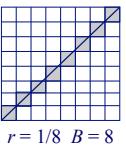
---



## Reminder: Hausdorff Dimension ( $D_0$ )



- $r$  = side length (each dimension)
- $B(r) = \# \text{ boxes containing points} \propto r^{D_0}$

 $r = 1/2$  $\log r = -1$   
 $\log B = 1$  $r = 1/4$  $\log r = -2$   
 $\log B = 2$  $r = 1/8$  $\log r = -3$   
 $\log B = 3$ 

15-826

Copyright: C. Faloutsos (2011)

29

---



---



---



---



---



---



---



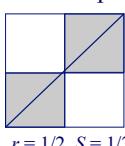
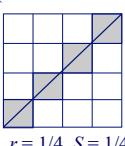
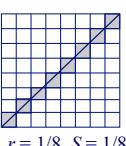
---



## Reminder: Correlation Dimension ( $D_2$ )



- $S(r) = \sum p_i^2$  (squared % pts in box)  $\propto r^{D_2}$   
 $\propto \#\text{pairs( within } \leq r)$

 $r = 1/2$  $\log r = -1$   
 $\log S = -1$  $r = 1/4$  $\log r = -2$   
 $\log S = -2$  $r = 1/8$  $\log r = -3$   
 $\log S = -3$ 

15-826

Copyright: C. Faloutsos (2011)

30

---



---



---



---



---



---



---



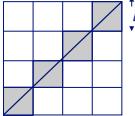
---

 CMU SCS

## Observation #1

proof

- How to determine avg MBR side  $l$ ?
  - $N = \# \text{pts}$ ,  $C = \text{MBR capacity}$



Hausdorff dimension:  $B(r) \propto r^{D_0}$

$$B(l) = N/C = l^{-D_0} \Rightarrow l = (N/C)^{-1/D_0}$$

15-826 Copyright: C. Faloutsos (2011) 31

---

---

---

---

---

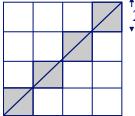
---

 CMU SCS

## Observation #2

proof

- $k$ -NN query  $\rightarrow \varepsilon$ -range query
  - For  $k$  pts, what radius  $\varepsilon$  do we expect?



Correlation dimension:  $S(r) \propto r^{D_2}$

$$S(\varepsilon) = \frac{k}{N-1} = (2\varepsilon)^{D_2}$$

15-826 Copyright: C. Faloutsos (2011) 32

---

---

---

---

---

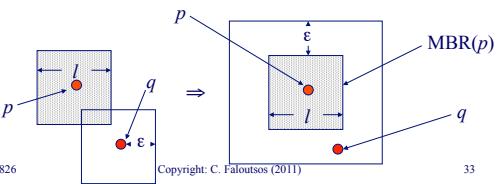
---

 CMU SCS

## Observation #3

proof

- Estimate avg # query-sensitive anchors:
  - How many **expected**  $q$  will touch **avg** page?
  - Page touch:  $q$  stabs  $\varepsilon$ -dilated MBR( $p$ )



15-826 Copyright: C. Faloutsos (2011) 33

---

---

---

---

---

---



## Asymptotic Formula

- $k$ -NN page accesses as  $N \rightarrow \infty$ 
  - $C$  = page capacity
  - $D$  = fractal dimension ( $=D_0 \sim D_2$ )

$$P_{all}^{L_\infty}(k) \approx \sum_{j=0}^h \left\{ \frac{1}{C^{h-j}} + \left[ 1 + \left( \frac{k}{C^{h-j}} \right)^{1/D} \right]^D \right\}$$

15-826

Copyright: C. Faloutsos (2011)

34

---



---



---



---



---



---



---



---



---



---



## Asymptotic Formula

$$P_{all}^{L_\infty}(k) \approx \sum_{j=0}^h \left\{ \frac{1}{C^{h-j}} + \left[ 1 + \left( \frac{k}{C^{h-j}} \right)^{1/D} \right]^D \right\}$$

- NO mention of the embedding dimensionality!!
- Still have dim. curse, but on f.d.  $D$

15-826

Copyright: C. Faloutsos (2011)

35

---



---



---



---



---



---



---



---



---

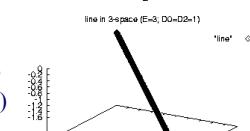


---



## Synthetic Data

- plane
  - $D_0 = D_2 = 2$
  - embedded in  $E$ -space
  - $N = 100K$
- manifold
  - $E = 8$
  - $D_0 = D_2$  varies from 1-6
  - line, plane, etc. (in 8-d)



15-826

Copyright: C. Faloutsos

---



---



---



---



---



---



---



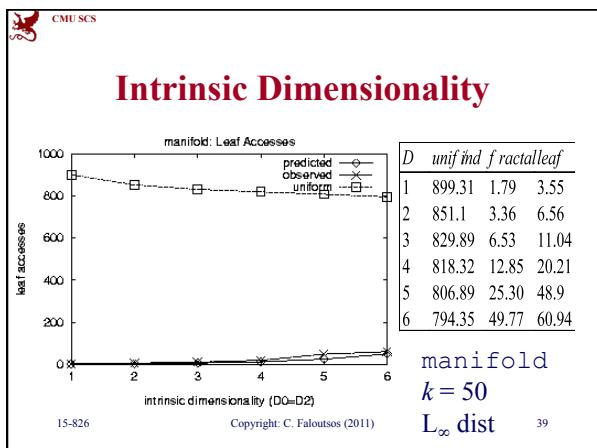
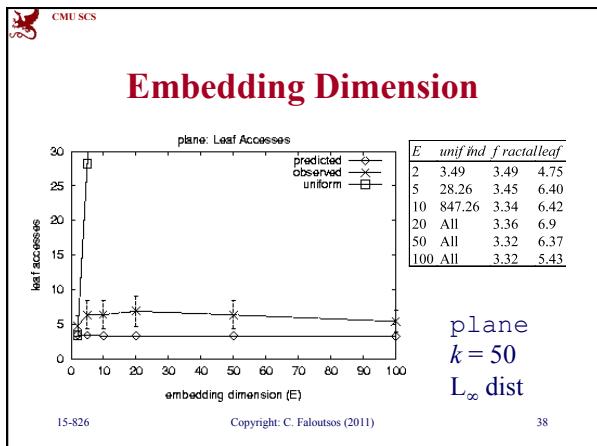
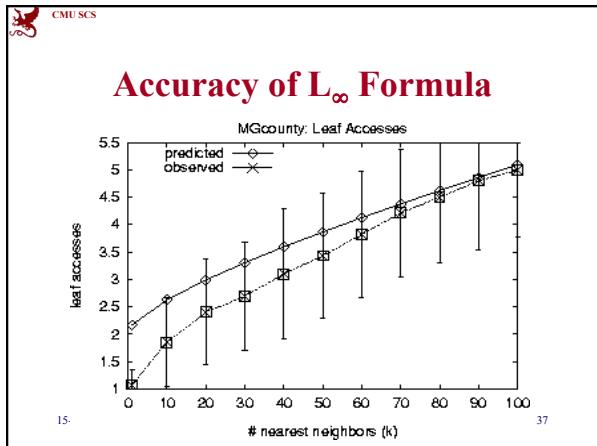
---



---



---





## Non-Euclidean Data Set

$E$	<i>uniform</i>	<i>fractal</i>	<i>leaf</i>
2	3.49	2.53	$4.72 \pm 1.81$
10	847.26	2.53	$6.42 \pm 2.11$
20	all	2.53	$7.76 \pm 4.12$
50	all	2.53	$6.15 \pm 2.82$
100	all	2.53	$5.64 \pm 2.32$

15-826

sierpinski,  $k = 50$ ,  $L_\infty$  dist

40

---



---



---



---



---



---



---



---



---



## Conclusions

- Worst-case theory is **over-pessimistic**
- High dimensional data can exhibit good performance if **correlated, non-uniform**
- Many real data sets are **self-similar**
- Determinant is **intrinsic** dimensionality
  - multiple fractal dimensions ( $D_0$  and  $D_2$ )
  - indication of how far one can go

15-826

Copyright: C. Faloutsos (2011)

41

---



---



---



---



---



---



---



---



---



## References

- Ciaccia, P., M. Patella, et al. (1998). *A Cost Model for Similarity Queries in Metric Spaces*. PODS.
- Pagel, B.-U., F. Korn, et al. (2000). *Deflating the Dimensionality Curse Using Multiple Fractal Dimensions*. ICDE, San Diego, CA.
- Traina, C., A. J. M. Traina, et al. (2000). *Distance Exponent: A New Concept for Selectivity Estimation in Metric Trees*. ICDE, San Diego, CA.

15-826

Copyright: C. Faloutsos (2011)

42

---



---



---



---



---



---



---



---



---