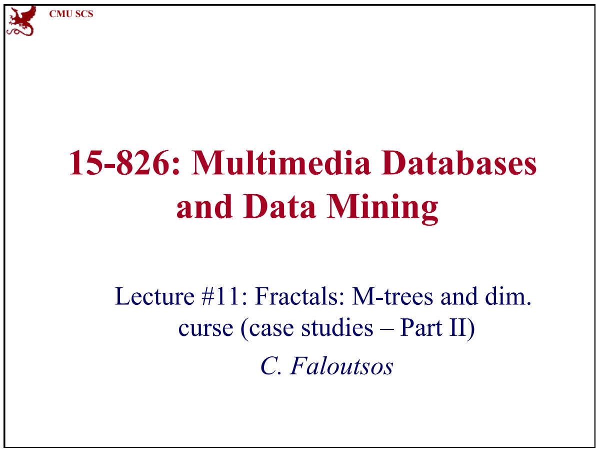
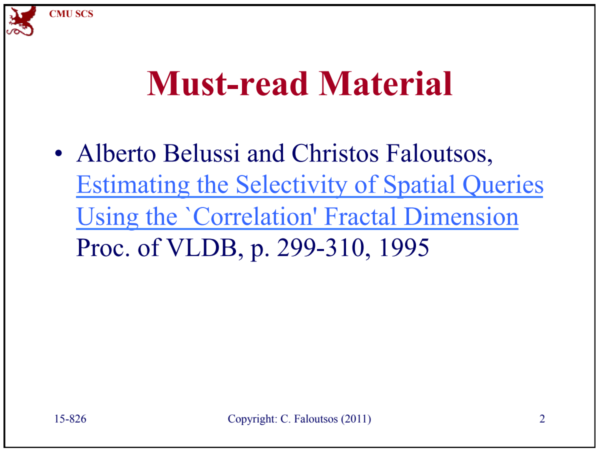


CMU SCS

15-826: Multimedia Databases and Data Mining

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

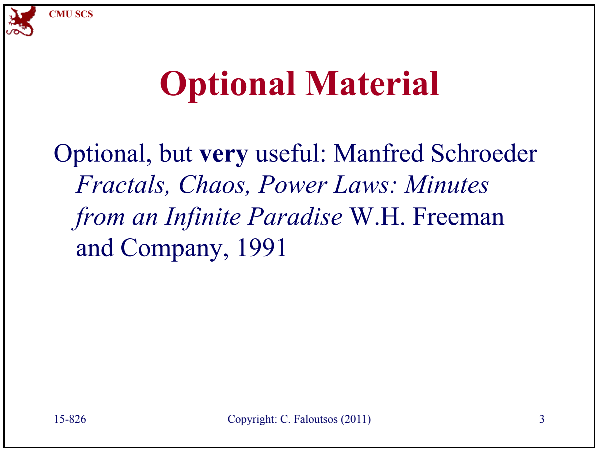
C. Faloutsos



CMU SCS

Must-read Material

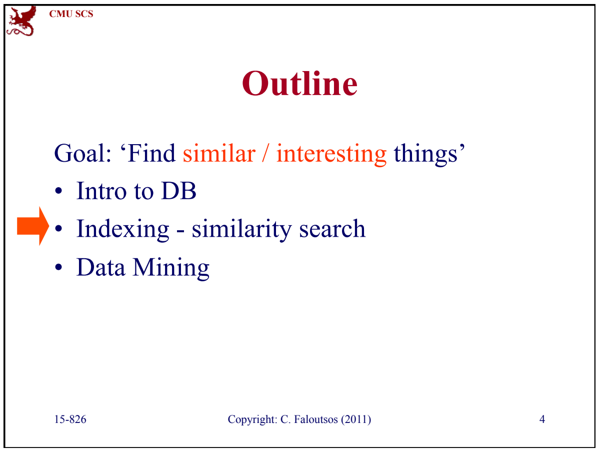
- Alberto Belussi and Christos Faloutsos, Estimating the Selectivity of Spatial Queries Using the `Correlation' Fractal Dimension Proc. of VLDB, p. 299-310, 1995



CMU SCS

Optional Material

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

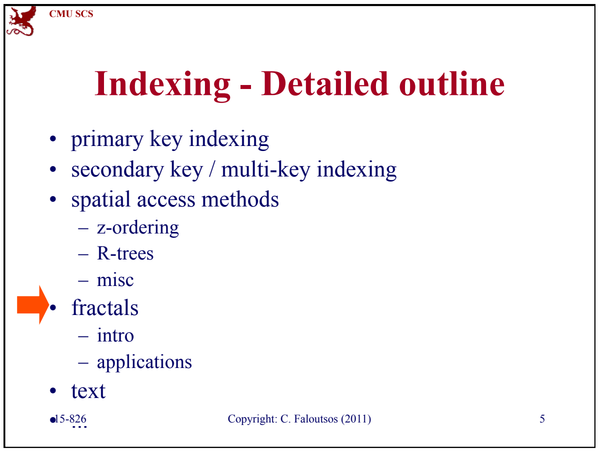


CMU SCS

Outline

Goal: 'Find similar / interesting things'

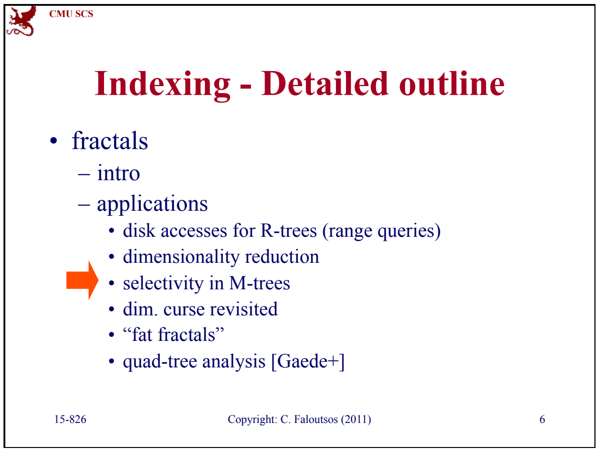
- Intro to DB
- ➡ Indexing - similarity search
- Data Mining



CMU SCS

Indexing - Detailed outline

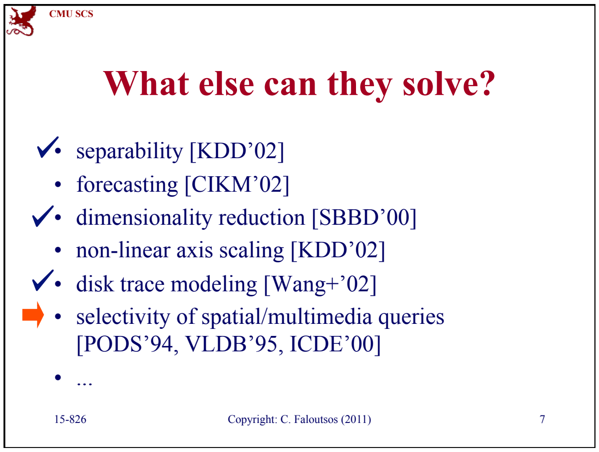
- primary key indexing
- secondary key / multi-key indexing
- spatial access methods
  - z-ordering
  - R-trees
  - misc
- ➡ fractals
  - intro
  - applications
- text



CMU SCS

Indexing - Detailed outline

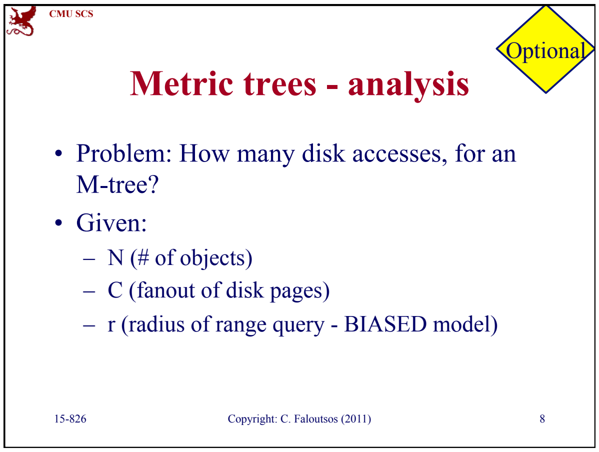
- fractals
  - intro
  - applications
    - disk accesses for R-trees (range queries)
    - dimensionality reduction
    - ➡ selectivity in M-trees
    - dim. curse revisited
    - "fat fractals"
    - quad-tree analysis [Gaede+]



CMU SCS

What else can they solve?

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

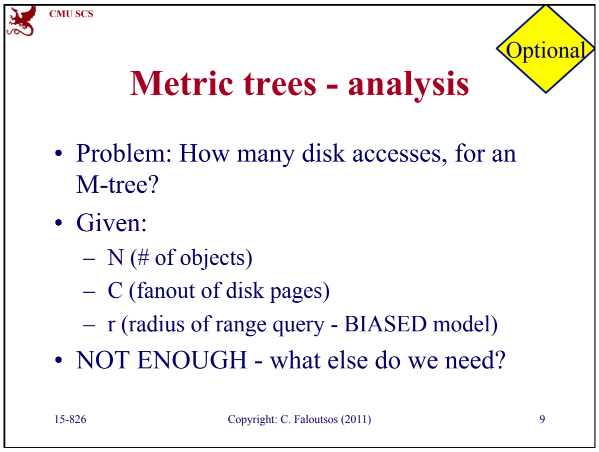


CMU SCS

Optional

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)

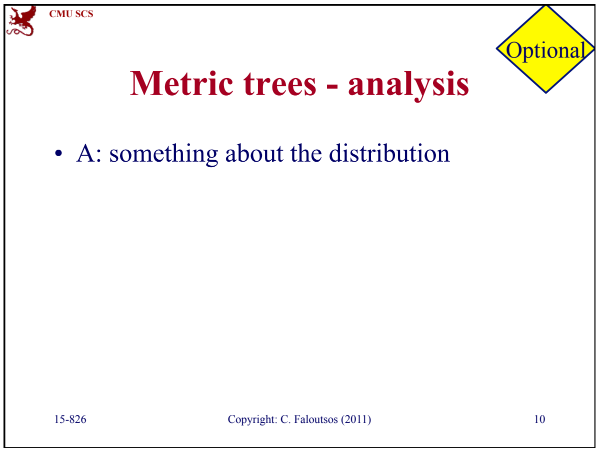


CMU SCS

Optional

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?

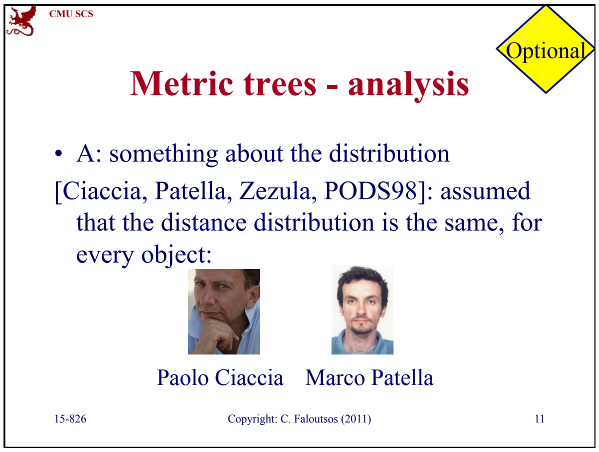


CMU SCS

Optional

Metric trees - analysis

- A: something about the distribution



CMU SCS

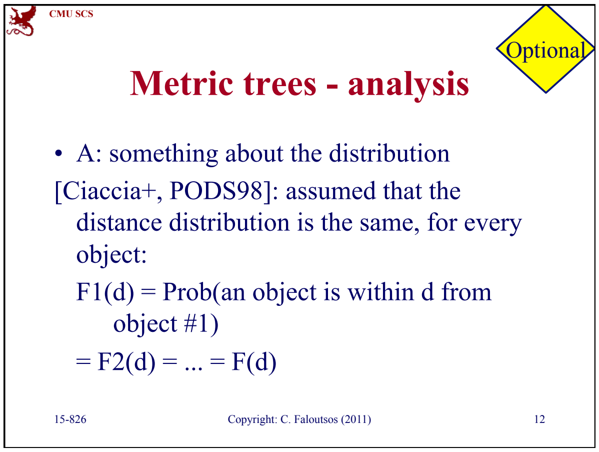
Optional

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



CMU SCS

Optional

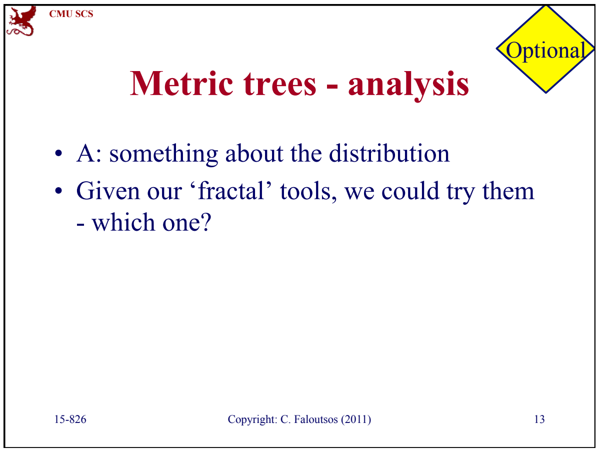
Metric trees - analysis

- A: something about the distribution

[Ciaccia+, PODS98]: assumed that the distance distribution is the same, for every object:

F1(d) = Prob(an object is within d from object #1)

= F2(d) = ... = F(d)



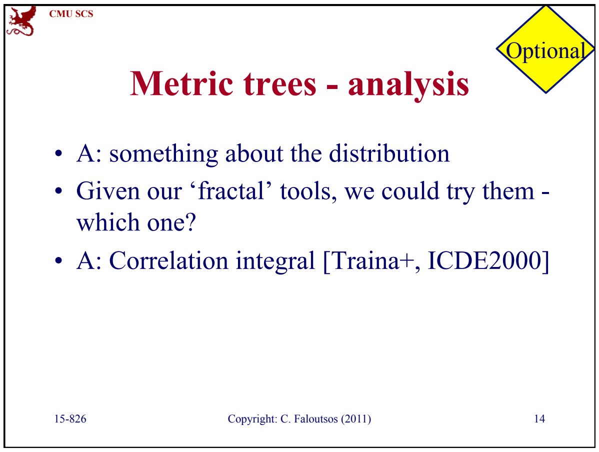
CMU SCS

Optional

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



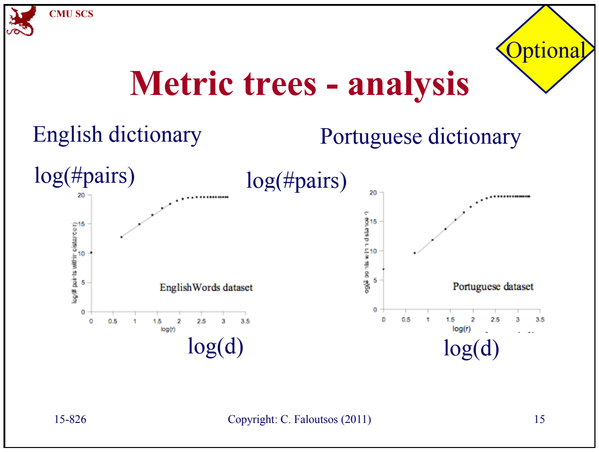
CMU SCS

Optional

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



CMU SCS

Optional

Metric trees - analysis

English dictionary

Portuguese dictionary

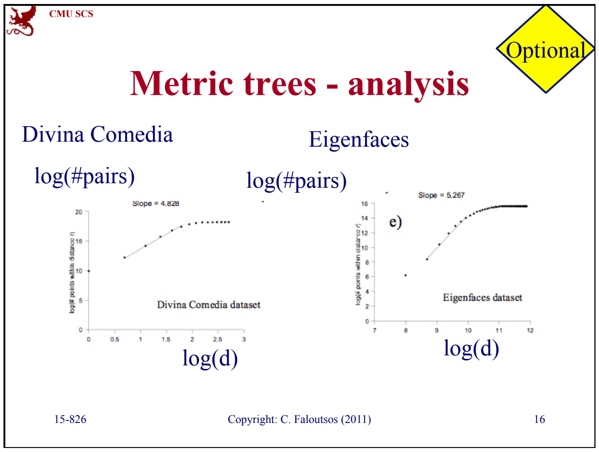
log(#pairs)

log(#pairs)

log(d)

log(d)

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



CMU SCS

Optional

Metric trees - analysis

Divina Comedia

Eigenfaces

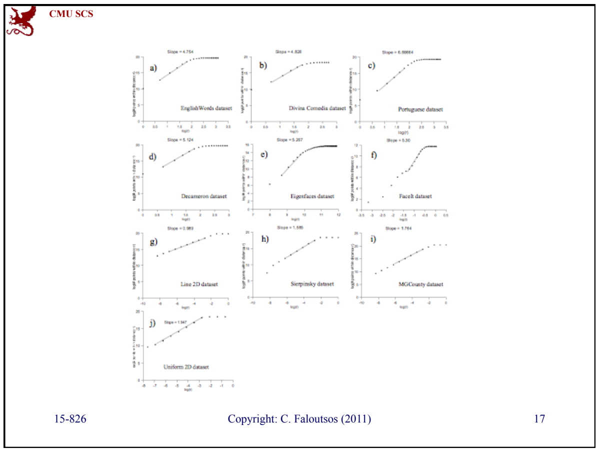
log(#pairs)

log(#pairs)

log(d)

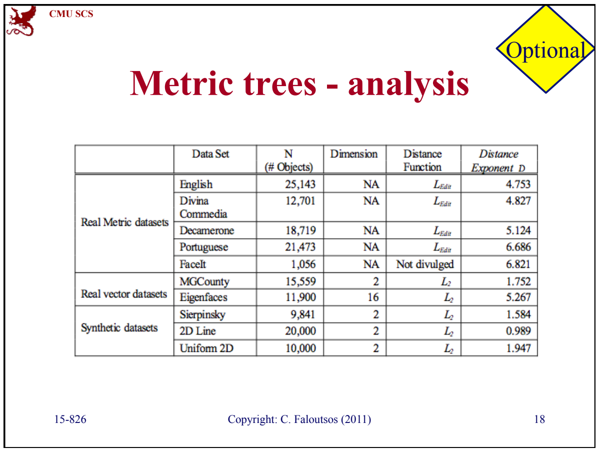
log(d)

15-826 Copyright: C. Faloutsos (2011) 16



CMU SCS

15-826 Copyright: C. Faloutsos (2011) 17



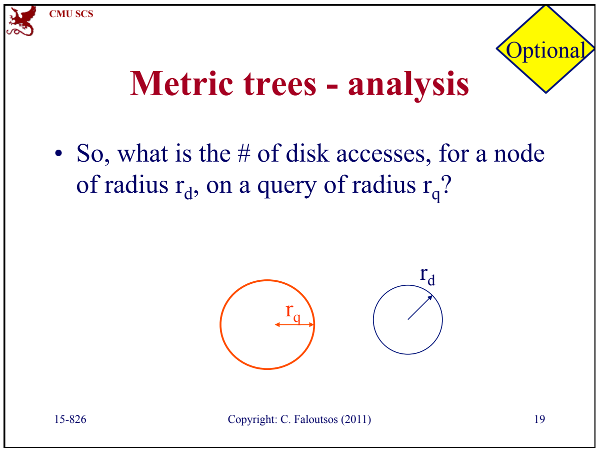
CMU SCS

Optional

Metric trees - analysis

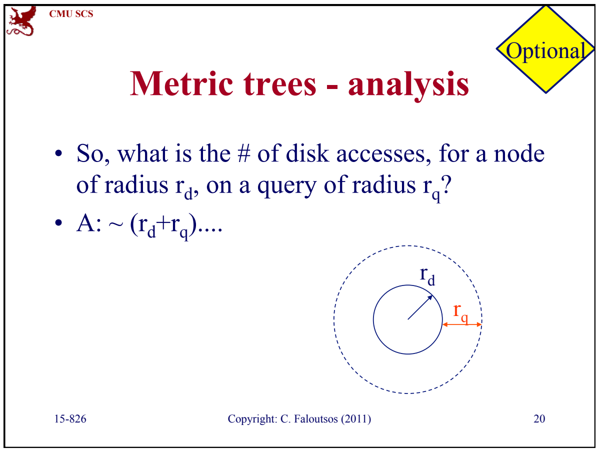
	Data Set	N (# Objects)	Dimension	Distance Function	*Distance Exponent D*
Real Metric datasets	English	25,143	NA	L_{Edit}	4.753
	Divina Commedia	12,701	NA	L_{Edit}	4.827
	Decamerone	18,719	NA	L_{Edit}	5.124
	Portuguese	21,473	NA	L_{Edit}	6.686
	FaceIt	1,056	NA	Not divulged	6.821
Real vector datasets	MGCounty	15,559	2	L_2	1.752
	Eigenfaces	11,900	16	L_2	5.267
Synthetic datasets	Sierpinsky	9,841	2	L_2	1.584
	2D Line	20,000	2	L_2	0.989
	Uniform 2D	10,000	2	L_2	1.947

15-826 Copyright: C. Faloutsos (2011) 18



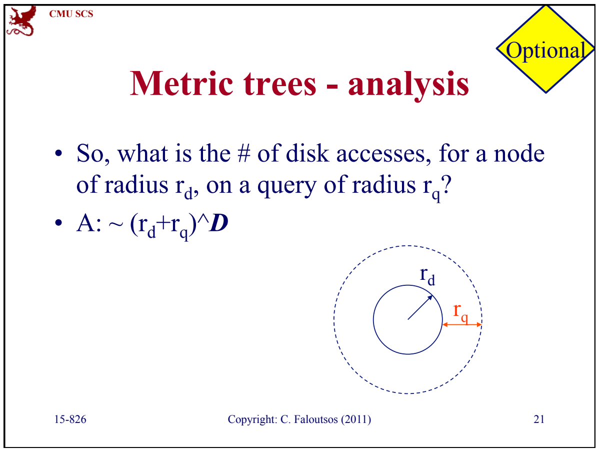
Metric trees - analysis

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



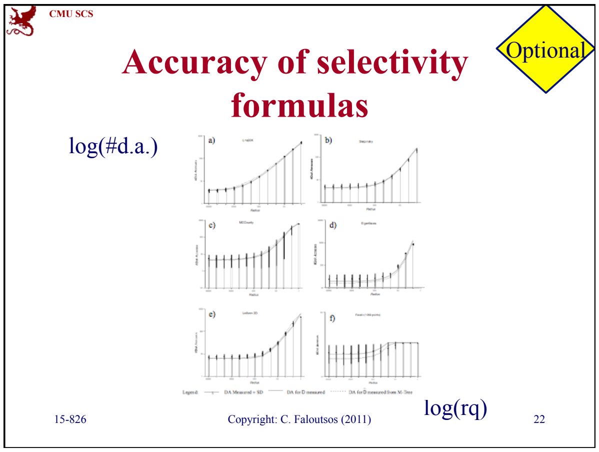
Metric trees - analysis

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



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: ~ (r_d+r_q)^***D***



CMU SCS

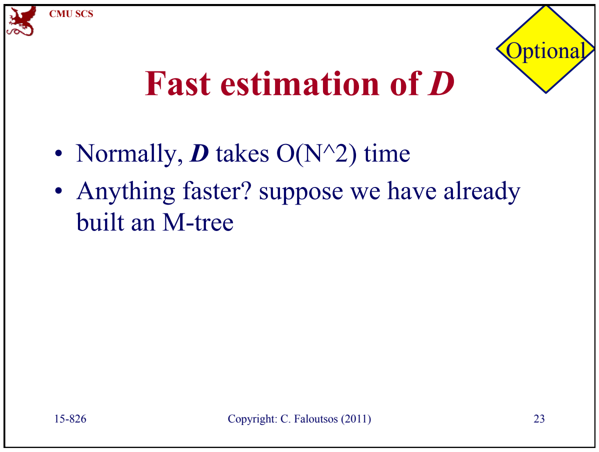
Optional

Accuracy of selectivity formulas

log(#d.a.)

log(rq)

15-826 22



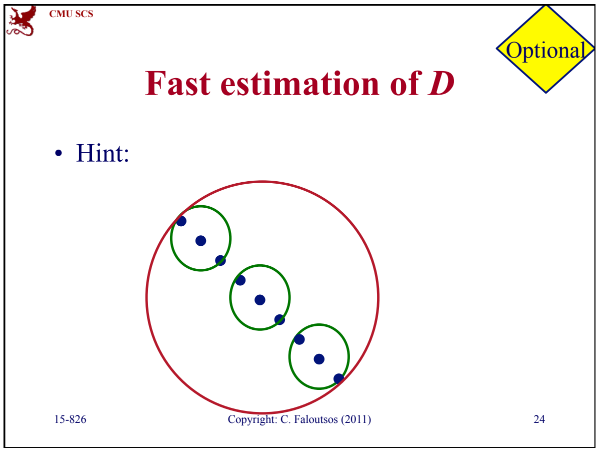
CMU SCS

Optional

Fast estimation of *D*

- Normally, ***D*** takes O(N^2) time
- Anything faster? suppose we have already built an M-tree

15-826 23



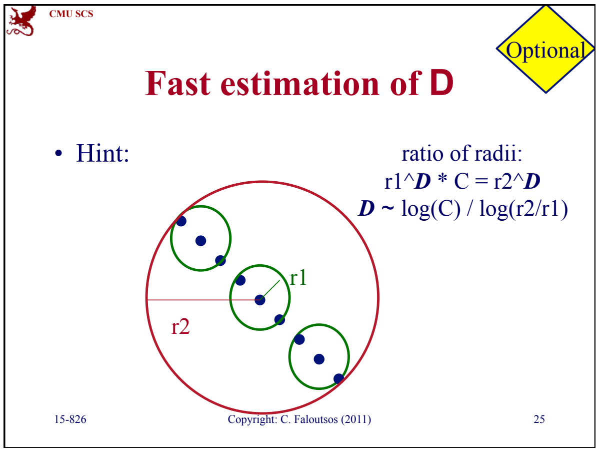
CMU SCS

Optional

Fast estimation of *D*

- Hint:

15-826 24



CMU SCS

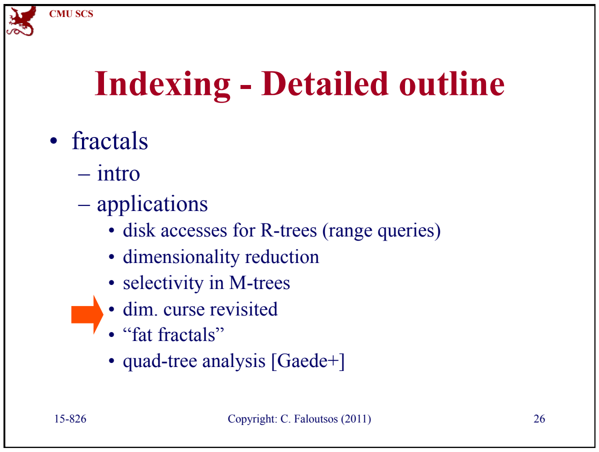
Optional

Fast estimation of D

- Hint:

ratio of radii:
r1^***D*** * C = r2^***D***
D ~ log(C) / log(r2/r1)

15-826 Copyright: C. Faloutsos (2011) 25

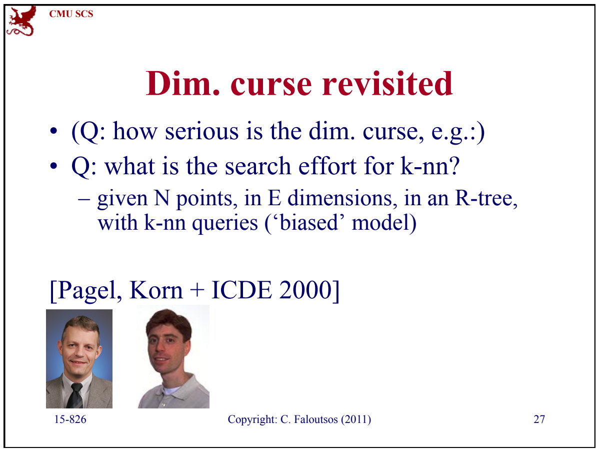


CMU SCS

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



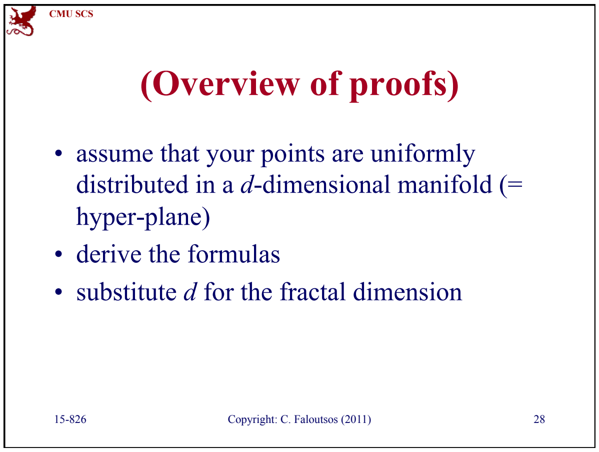
CMU SCS

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)

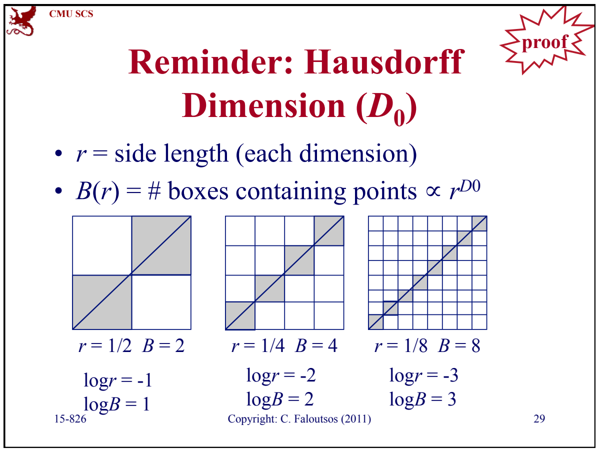
[Pagel, Korn + ICDE 2000]

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



proof

Reminder: Hausdorff Dimension (D_0)

- r = side length (each dimension)
- $B(r)$ = # boxes containing points $\propto r^{D0}$

$r = 1/2 \quad B = 2$

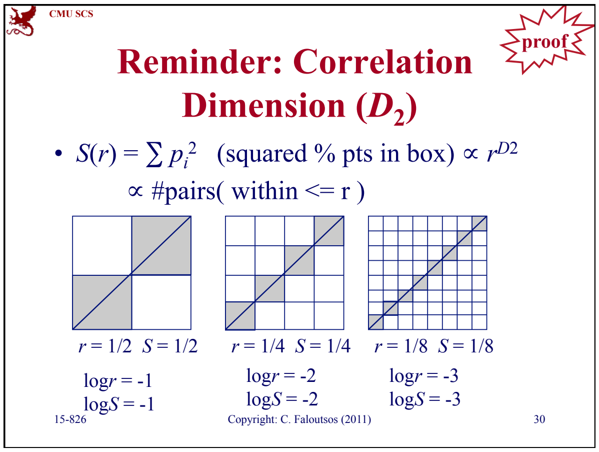
$\log r = -1$
$\log B = 1$

$r = 1/4 \quad B = 4$

$\log r = -2$
$\log B = 2$

$r = 1/8 \quad B = 8$

$\log r = -3$
$\log B = 3$



proof

Reminder: Correlation Dimension (D_2)

- $S(r) = \sum p_i^2$ (squared % pts in box) $\propto r^{D2}$

  $\propto$ #pairs(within <= r)

$r = 1/2 \quad S = 1/2$

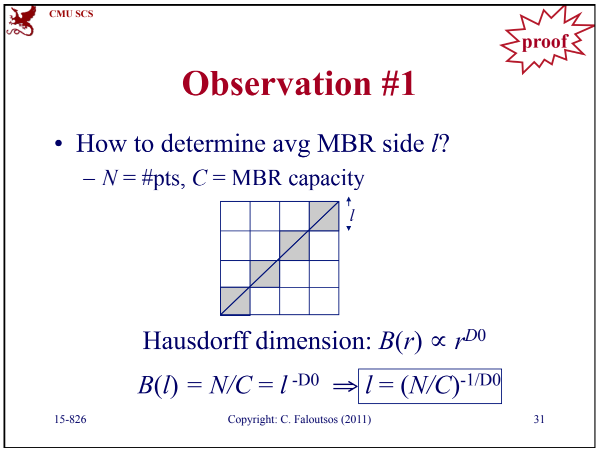
$\log r = -1$
$\log S = -1$

$r = 1/4 \quad S = 1/4$

$\log r = -2$
$\log S = -2$

$r = 1/8 \quad S = 1/8$

$\log r = -3$
$\log S = -3$



CMU SCS

proof

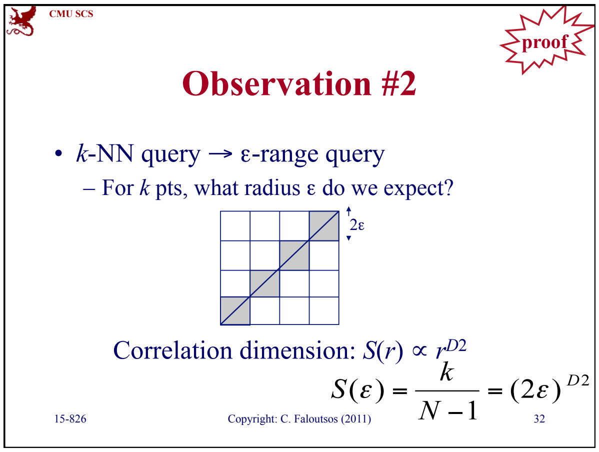
Observation #1

- How to determine avg MBR side l?
  - N = #pts, C = MBR capacity

Hausdorff dimension: $B(r) \propto r^{D0}$

$B(l) = N/C = l^{-D0} \Rightarrow \boxed{l = (N/C)^{-1/D0}}$

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



CMU SCS

proof

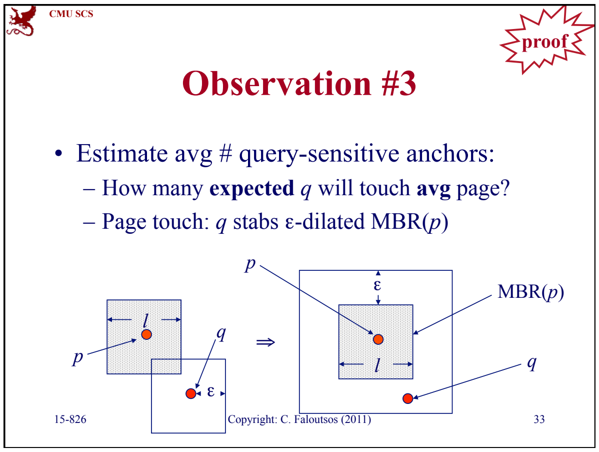
Observation #2

- k-NN query → ε-range query
  - For k pts, what radius ε do we expect?

Correlation dimension: $S(r) \propto r^{D2}$

$$S(\varepsilon) = \frac{k}{N-1} = (2\varepsilon)^{D2}$$

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



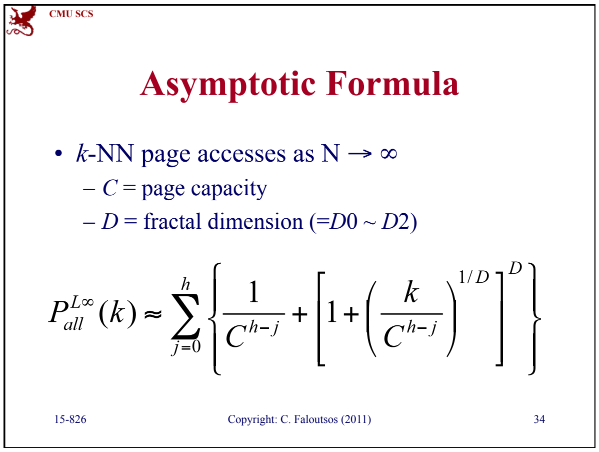
CMU SCS

proof

Observation #3

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

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

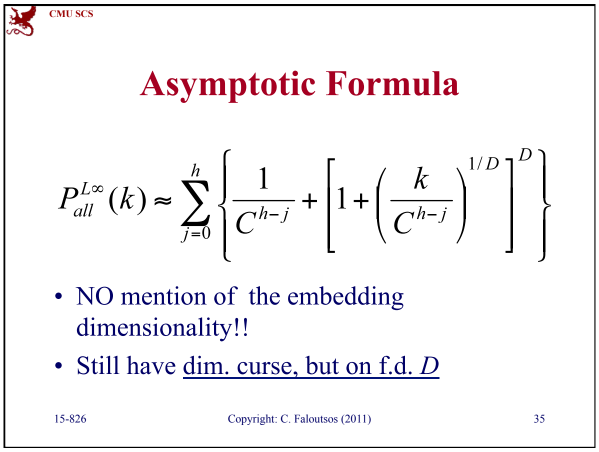


CMU SCS

Asymptotic Formula

- *k*-NN page accesses as N → ∞
  - *C* = page capacity
  - *D* = fractal dimension (=*D*0 ~ *D*2)

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

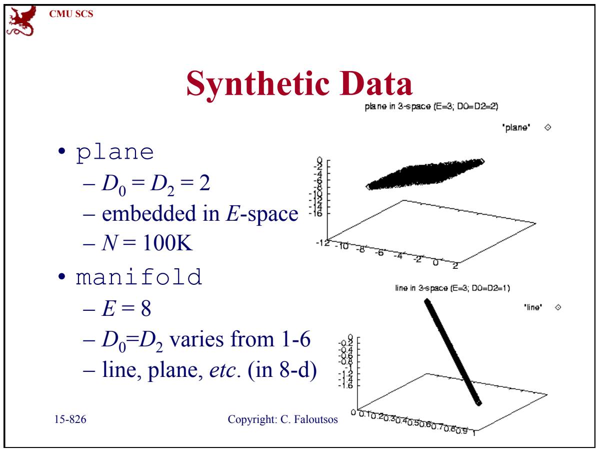


CMU SCS

Asymptotic Formula

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

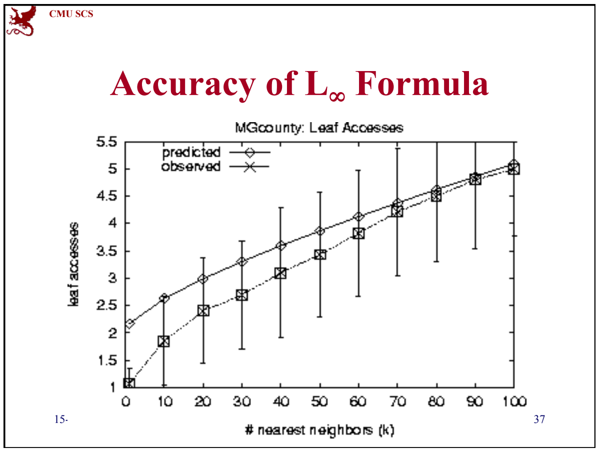
- NO mention of the embedding dimensionality!!
- Still have dim. curse, but on f.d. *D*



CMU SCS

Synthetic Data

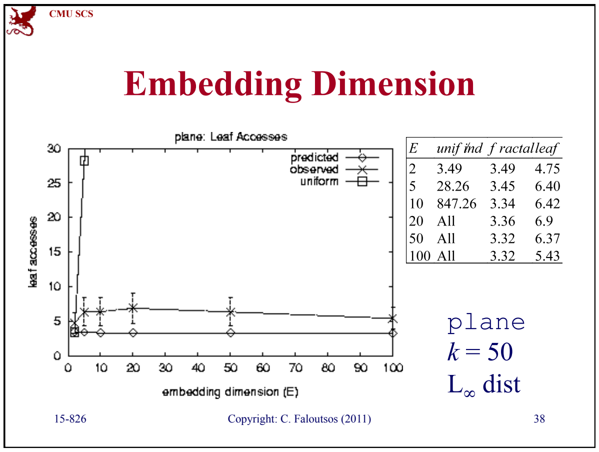
- plane
  - $D_0 = D_2 = 2$
  - embedded in *E*-space
  - *N* = 100K
- manifold
  - *E* = 8
  - D_0=D_2 varies from 1-6
  - line, plane, *etc*. (in 8-d)



CMU SCS

Accuracy of L_∞ Formula

15- 37



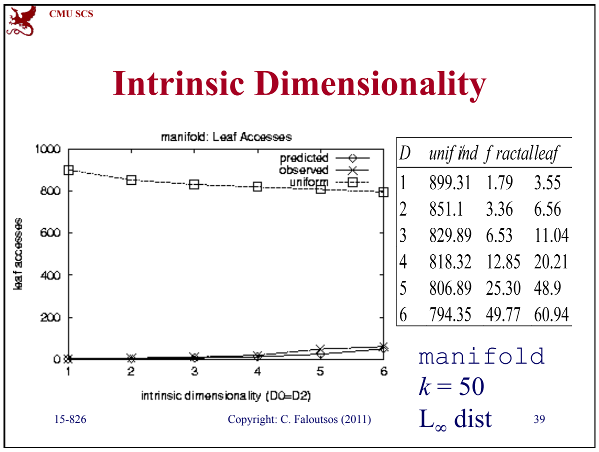
CMU SCS

Embedding Dimension

E	*unif ind*	*f ractal*	*leaf*
2	3.49	3.49	4.75
5	28.26	3.45	6.40
10	847.26	3.34	6.42
20	All	3.36	6.9
50	All	3.32	6.37
100	All	3.32	5.43

`plane`
$k = 50$
L_∞ dist

15-826 Copyright: C. Faloutsos (2011) 38



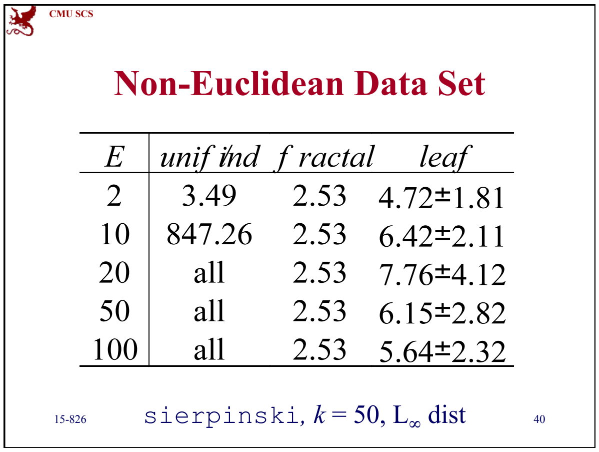
CMU SCS

Intrinsic Dimensionality

D	*unif ind*	*f ractal*	*leaf*
1	899.31	1.79	3.55
2	851.1	3.36	6.56
3	829.89	6.53	11.04
4	818.32	12.85	20.21
5	806.89	25.30	48.9
6	794.35	49.77	60.94

`manifold`
$k = 50$
L_∞ dist

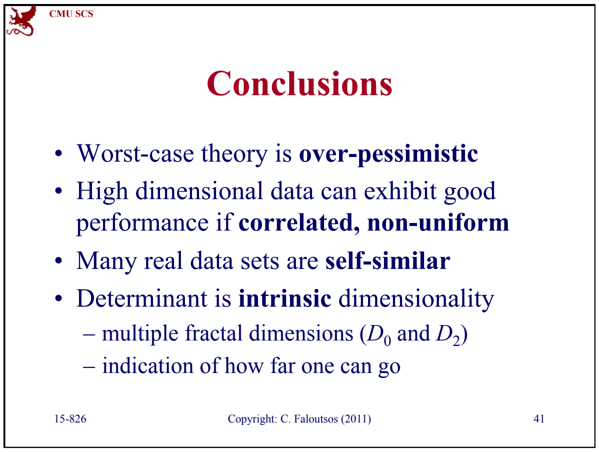
15-826 Copyright: C. Faloutsos (2011) 39



Non-Euclidean Data Set

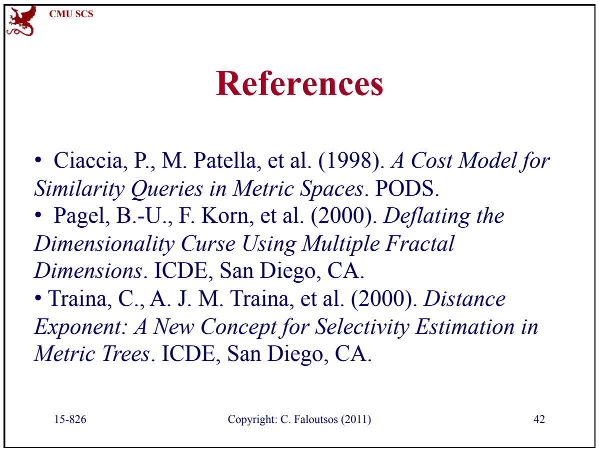
E	*unif ind*	*f ractal*	*leaf*
2	3.49	2.53	4.72±1.81
10	847.26	2.53	6.42±2.11
20	all	2.53	7.76±4.12
50	all	2.53	6.15±2.82
100	all	2.53	5.64±2.32

`sierpinski`, $k = 50$, L_∞ dist



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



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.