


**15-826: Multimedia Databases
and Data Mining**


Lecture #9: Fractals - introduction
C. Faloutsos



Must-read Material

- Christos Faloutsos and Ibrahim Kamel, *Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension*, Proc. ACM SIGACT-SIGMOD-SIGART PODS, May 1994, pp. 4-13, Minneapolis, MN.

15-826 Copyright: C. Faloutsos (2009) 2



Recommended Material

optional, but **very** useful:

- Manfred Schroeder *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* W.H. Freeman and Company, 1991 (on reserve in the library)
 - Chapter 10: boxcounting method
 - Chapter 1: Sierpinski triangle

15-826 Copyright: C. Faloutsos (2009) 3

CMU SCS

Outline

Goal: 'Find similar / interesting things'

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

15-826 Copyright: C. Faloutsos (2009) 4

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

15-826 Copyright: C. Faloutsos (2009) 5

CMU SCS

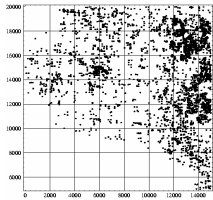
Intro to fractals - outline

- ➔ • Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2009) 6

CMU SCS

Problem #1: GIS - points



Road end-points of Montgomery county:

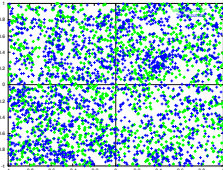
- Q1: how many d.a. for an R-tree?
- Q2 : distribution?
 - not uniform
 - not Gaussian
 - no rules??

15-826 Copyright: C. Faloutsos (2009) 7

CMU SCS

Problem #2 - spatial d.m.

Galaxies (Sloan Digital Sky Survey w/ B. Nichol)



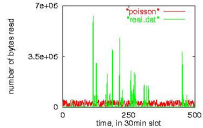
- 'spiral' and 'elliptical' galaxies
(stores and households ...)
- patterns?
- attraction/repulsion?
- how many 'spi' within r from an 'ell'?

15-826 Copyright: C. Faloutsos (2009) 8

CMU SCS

Problem #3: traffic

bytes



- disk trace (from HP - J. Wilkes); Web traffic - fit a model
- how many explosions to expect?
- queue length distr.?

15-826 Copyright: C. Faloutsos (2009) 9

CMU SCS

Problem #3: traffic

bytes

time

Poisson
indep.,
ident. distr

15-826 Copyright: C. Faloutsos (2009) 10

CMU SCS

Problem #3: traffic

bytes

time

~~Poisson~~
~~indep.,~~
~~ident. distr~~

15-826 Copyright: C. Faloutsos (2009) 11

CMU SCS

Problem #3: traffic

bytes

time

~~Poisson~~
~~indep.,~~
~~ident. distr~~

Q: Then, how to generate such bursty traffic?

15-826 Copyright: C. Faloutsos (2009) 12

CMU SCS

Common answer:

- Fractals / self-similarities / power laws
- Seminal works from Hilbert, Minkowski, Cantor, Mandelbrot, (Hausdorff, Lyapunov, Ken Wilson, ...)

15-826 Copyright: C. Faloutsos (2009) 13

CMU SCS

Road map

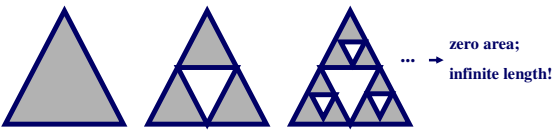
- Motivation – 3 problems / case studies
- ➔ • Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2009) 14

CMU SCS

What is a fractal?

= self-similar point set, e.g., Sierpinski triangle:



... → zero area; infinite length!

15-826 Copyright: C. Faloutsos (2009) 15

CMU SCS

Definitions (cont'd)

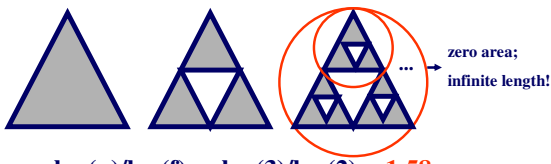
- Paradox: Infinite perimeter ; Zero area!
- 'dimensionality': between 1 and 2
- actually: $\text{Log}(3)/\text{Log}(2) = 1.58\dots$

15-826 Copyright: C. Faloutsos (2009) 16

CMU SCS

Dfn of fd:

ONLY for a perfectly self-similar point set:



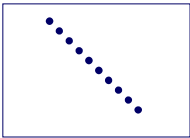
$=\log(n)/\log(f) = \log(3)/\log(2) = 1.58$

15-826 Copyright: C. Faloutsos (2009) 17

CMU SCS

Intrinsic ('fractal') dimension

- Q: fractal dimension of a line?
- A: 1 ($= \log(2)/\log(2)$!)

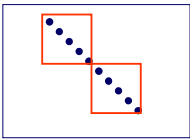


15-826 Copyright: C. Faloutsos (2009) 18

CMU SCS

Intrinsic ('fractal') dimension

- Q: fractal dimension of a line?
- A: 1 ($= \log(2)/\log(2)$)

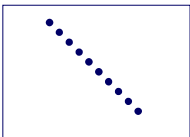


15-826 Copyright: C. Faloutsos (2009) 19

CMU SCS

Intrinsic ('fractal') dimension

- Q: dfn for a given set of points?



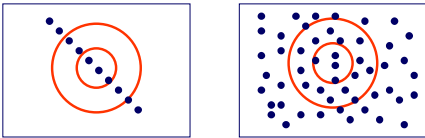
x	y
5	1
4	2
3	3
2	4

15-826 Copyright: C. Faloutsos (2009) 20

CMU SCS

Intrinsic ('fractal') dimension

- Q: fractal dimension of a line?
- A: $nn (\leq r) \sim r^1$ ('power law': $y=x^a$)
- Q: fd of a plane?
- A: $nn (\leq r) \sim r^2$
fd = slope of $(\log(nn) \text{ vs } \log(r))$



15-826 Copyright: C. Faloutsos (2009) 21

CMU SCS

Intrinsic ('fractal') dimension

- Algorithm, to estimate it?

Notice

- $avg\ nn(<=r)$ is exactly $tot\#pairs(<=r) / N$

including 'mirror' pairs

15-826 Copyright: C. Faloutsos (2009) 22

CMU SCS

Sierpinsky triangle

== 'correlation integral'

15-826 Copyright: C. Faloutsos (2009) 23

CMU SCS

Observations:


- Euclidean objects have **integer** fractal dimensions
 - point: 0
 - lines and smooth curves: 1
 - smooth surfaces: 2
- fractal dimension -> roughness of the periphery

15-826 Copyright: C. Faloutsos (2009) 24

CMU SCS

Important properties

- fd = embedding dimension -> uniform pointset
- a point set may have several fd, depending on scale




15-826 Copyright: C. Faloutsos (2009) 25

CMU SCS

Important properties

- fd = embedding dimension -> uniform pointset
- a point set may have several fd, depending on scale




2-d

15-826 Copyright: C. Faloutsos (2009) 26

CMU SCS

Important properties

- fd = embedding dimension -> uniform pointset
- a point set may have several fd, depending on scale

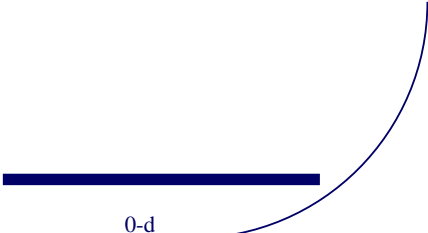


1-d

15-826 Copyright: C. Faloutsos (2009) 27

CMU SCS

Important properties



0-d

15-826 Copyright: C. Faloutsos (2009) 28

CMU SCS

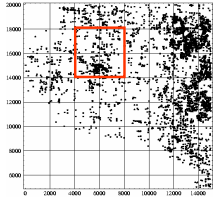
Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- ➔ • Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2009) 29

CMU SCS

Problem #1: GIS points



Cross-roads of
Montgomery county:

- any rules?

15-826 Copyright: C. Faloutsos (2009) 30

CMU SCS

Solution #1

$\log(\#\text{pairs}(\text{within } \leq r))$

SLOPE = 1.51847

1.51

$\log(r)$

A: self-similarity ->

- \Leftrightarrow fractals
- \Leftrightarrow scale-free
- \Leftrightarrow power-laws ($y=x^a, F=C*r^{(-2)}$)
- $\text{avg}\#\text{neighbors}(\leq r) = r^D$

15-826 Copyright: C. Faloutsos (2009) 31

CMU SCS

Solution #1

$\log(\#\text{pairs}(\text{within } \leq r))$

SLOPE = 1.51847

1.51

$\log(r)$

A: self-similarity

- $\text{avg}\#\text{neighbors}(\leq r) \sim r^{1.51}$

15-826 Copyright: C. Faloutsos (2009) 32

CMU SCS

Examples:MG county

- Montgomery County of MD (road endpoints)

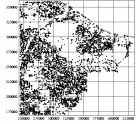
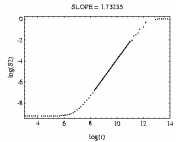
SLOPE = 1.51847

15-826 Copyright: C. Faloutsos (2009) 33

CMU SCS

Examples: LB county

- Long Beach county of CA (road end-points)

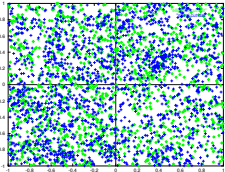
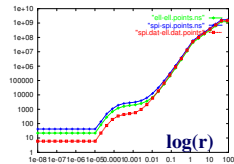



15-826 Copyright: C. Faloutsos (2009) 34

CMU SCS

Solution#2: spatial d.m.

Galaxies ('BOPS' plot - [sigmod2000])

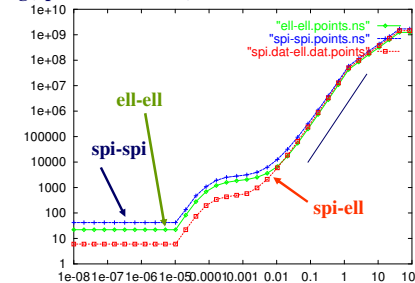



15-826 Copyright: C. Faloutsos (2009) 35

CMU SCS

Solution#2: spatial d.m.

log(#pairs within <=r)

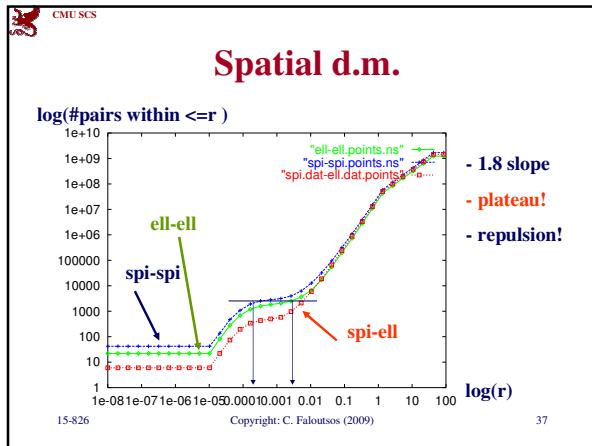


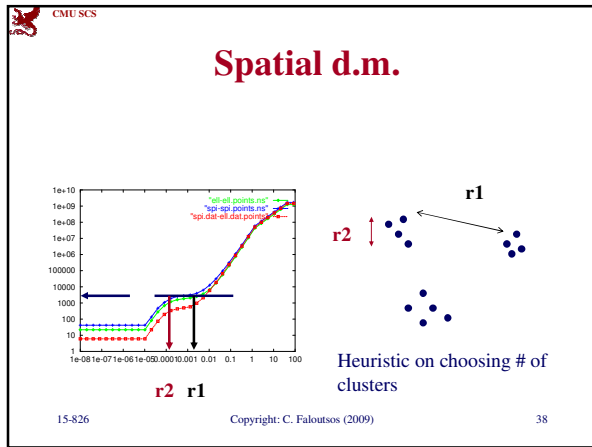
- 1.8 slope

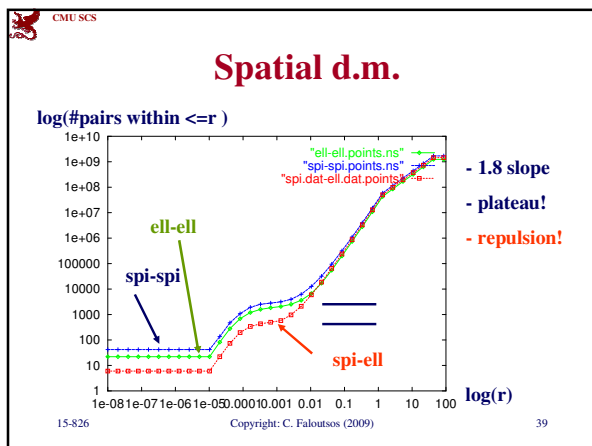
- plateau!

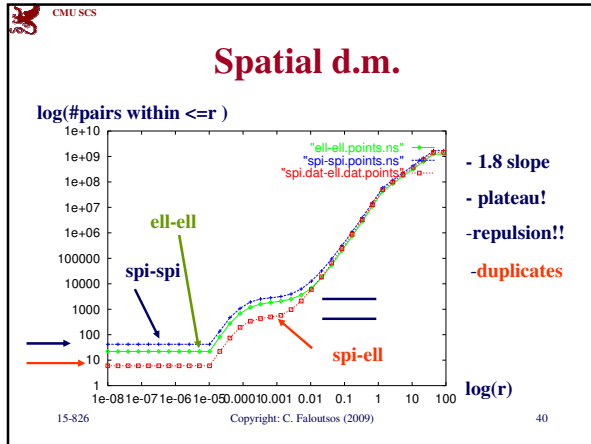
- repulsion!

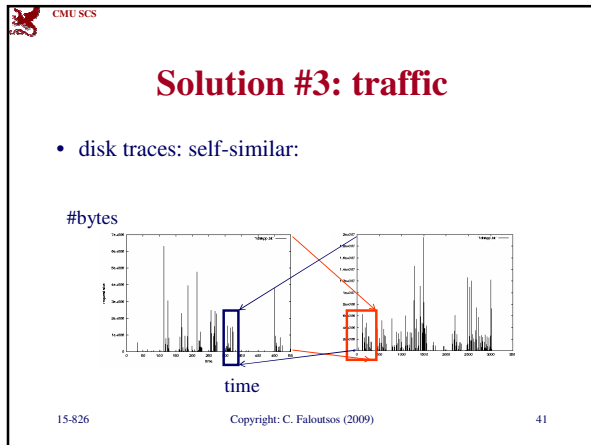
15-826 Copyright: C. Faloutsos (2009) 36

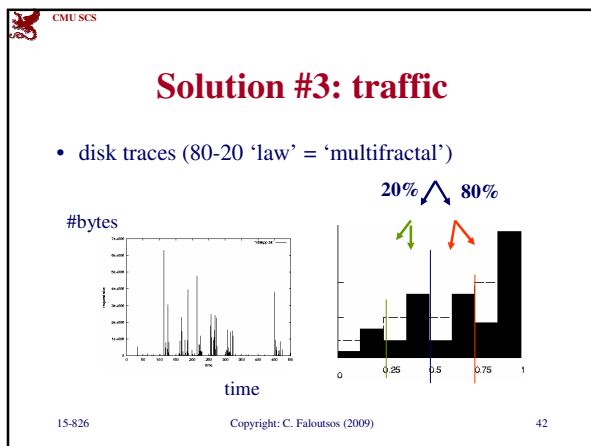


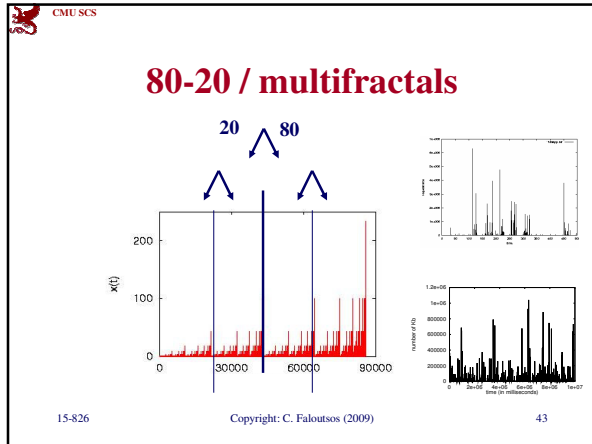


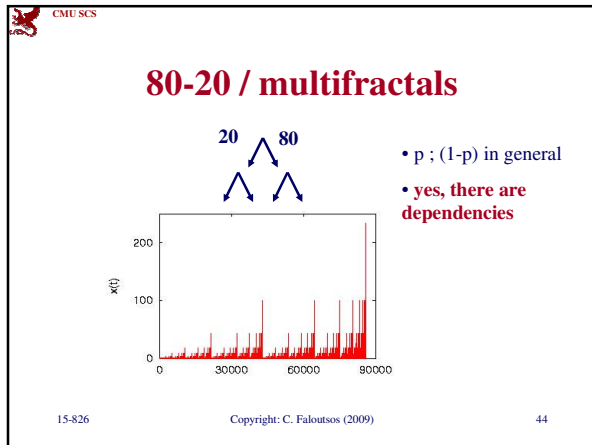


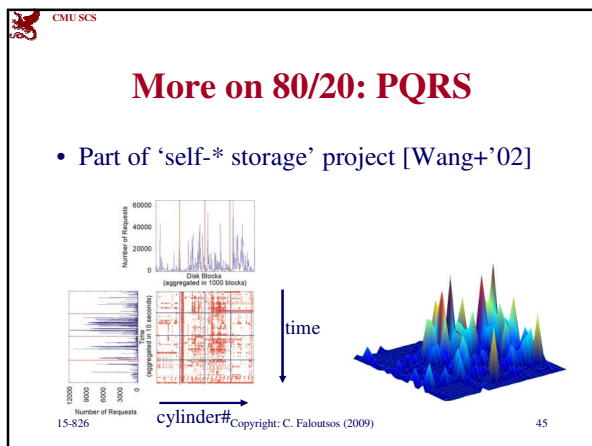












CMU SCS

More on 80/20: PQRS

- Part of 'self-* storage' project [Wang+'02]

15-826 Copyright: C. Faloutsos (2009) 46

CMU SCS

Solution#3: traffic

Clarification:

- fractal: a set of points that is self-similar
- multifractal: a probability density function that is self-similar

Many other time-sequences are bursty/clustered: (such as?)

15-826 Copyright: C. Faloutsos (2009) 47

CMU SCS

Example:

- network traffic

<http://repository.cs.vt.edu/lbl-conn-7.tar.Z>

15-826 Copyright: C. Faloutsos (2009) 48

CMU SCS

Web traffic

- [Crovella Bestavros, SIGMETRICS'96]

1000 sec; 100sec
10sec; 1sec

15-826 Copyright: C. Faloutsos (2009) 49

CMU SCS

Tape accesses

tapes needed, to retrieve n records?
(# days down, due to failures / hurricanes / communication noise...)

15-826 Copyright: C. Faloutsos (2009) 50

CMU SCS

Tape accesses

tapes retrieved

50-50 = Poisson
LD16

qual. records

15-826 Copyright: C. Faloutsos (2009) 51

CMU SCS

Road map

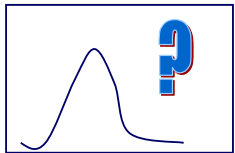
- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- ➔ More **tools** and examples
- Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2009) 52

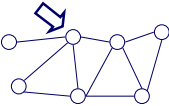
CMU SCS

A counter-intuitive example

count



- avg degree is, say 3.3
- pick a node at random – guess its degree, exactly (-> “mode”)

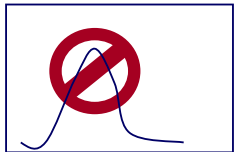


15-826 Copyright: C. Faloutsos (2009) 53

CMU SCS

A counter-intuitive example

count



- avg degree is, say 3.3
- pick a node at random – guess its degree, exactly (-> “mode”)
- A: 1!!

15-826 Copyright: C. Faloutsos (2009) 54

CMU SCS

A counter-intuitive example

count

avg: 3.3 degree

15-826 Copyright: C. Faloutsos (2009) 55

- avg degree is, say 3.3
- pick a node at random - what is the degree you expect it to have?
- A: 1!!
- A': very skewed distr.
- Corollary: **the mean is meaningless!**
- (and std -> infinity (!))

CMU SCS

Rank exponent R

- Power law in the degree distribution [SIGCOMM99]

internet domains

att.com
ibm.com

-0.82

15-826 Copyright: C. Faloutsos (2009) 56

CMU SCS

More tools

- Zipf's law
- Korcak's law / "fat fractals"

15-826 Copyright: C. Faloutsos (2009) 57

CMU SCS

A famous power law: Zipf's law

- Q: vocabulary word frequency in a document - any pattern?

aaron
ZOO

15-826
Copyright: C. Faloutsos (2009)
58

CMU SCS

A famous power law: Zipf's law

log(rank)

- Bible - rank vs frequency (log-log)

15-826
Copyright: C. Faloutsos (2009)
59

CMU SCS

A famous power law: Zipf's law

log(rank)

- Bible - rank vs frequency (log-log)
- similarly, in **many other** languages; for customers and sales volume; city populations etc etc

15-826
Copyright: C. Faloutsos (2009)
60

CMU SCS

A famous power law: Zipf's law

log(freq)

log(rank)

- Zipf distr:
 $freq = 1 / rank$
- generalized Zipf:
 $freq = 1 / (rank)^a$

15-826 Copyright: C. Faloutsos (2009) 61

CMU SCS

Olympic medals (Sidney):

log(#medals)

rank

$y = -0.9676x + 2.3241$
 $R^2 = 0.9458$

15-826 Copyright: C. Faloutsos (2009) 62

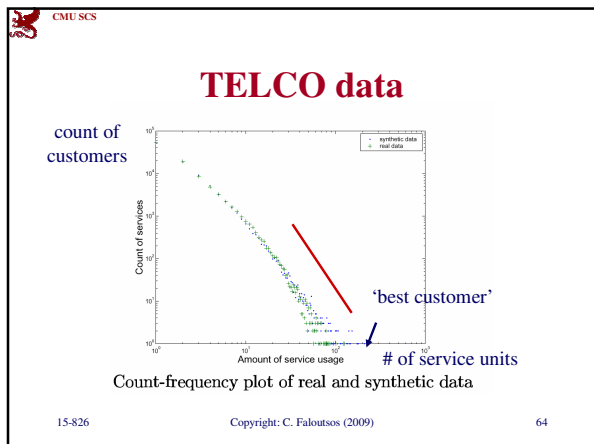
CMU SCS

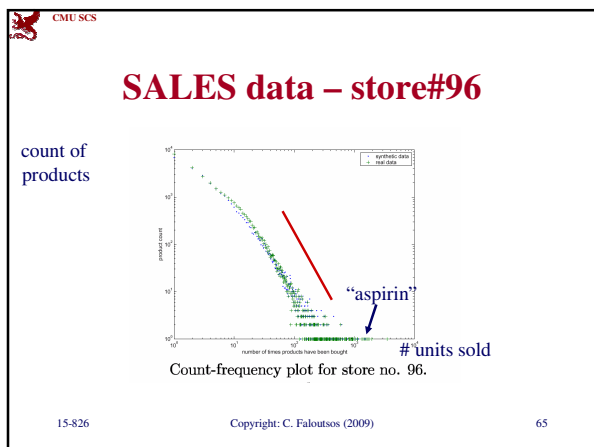
Olympic medals (Sidney'00, Athens'04):

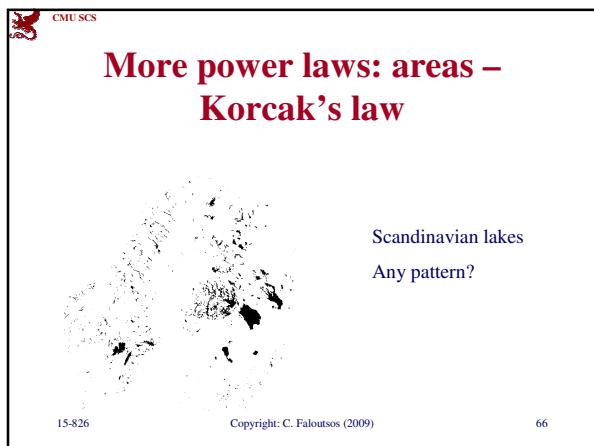
log(#medals)

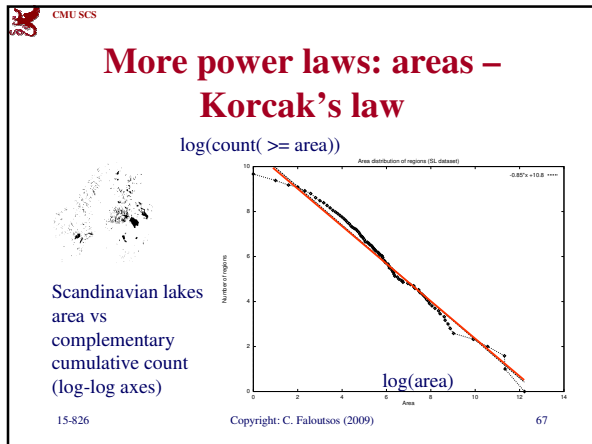
log(rank)

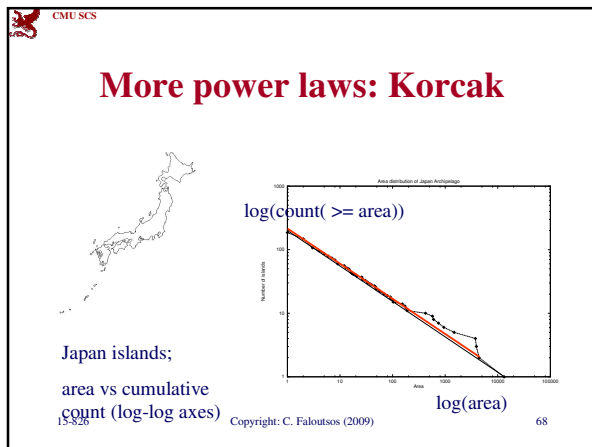
15-826 Copyright: C. Faloutsos (2009) 63

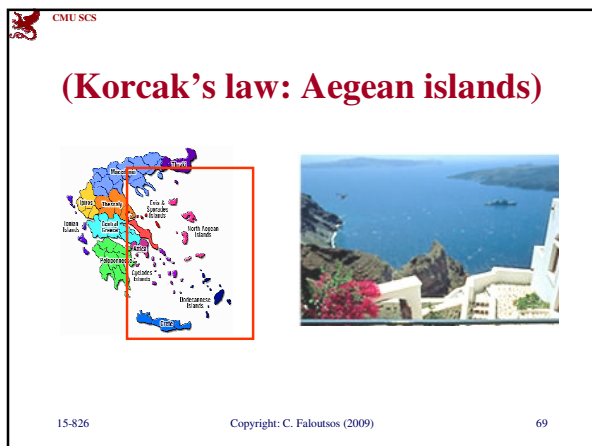













CMU SCS

Korcak's law & "fat fractals"



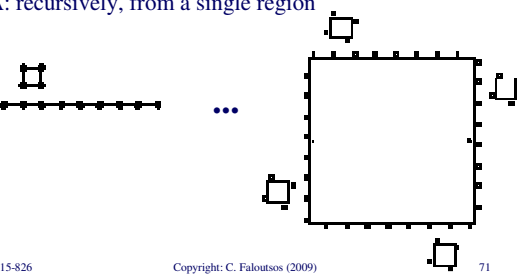
How to generate such regions?

15-826 Copyright: C. Faloutsos (2009) 70

CMU SCS

Korcak's law & "fat fractals"

Q: How to generate such regions?
A: recursively, from a single region



15-826 Copyright: C. Faloutsos (2009) 71

CMU SCS

so far we've seen:

- concepts:
 - fractals, multifractals and fat fractals
- tools:
 - correlation integral (= pair-count plot)
 - rank/frequency plot (Zipf's law)
 - CCDF (Korcak's law)

15-826 Copyright: C. Faloutsos (2009) 72

CMU SCS

so far we've seen:

- concepts:
 - fractals, multifractals and fat fractals
- tools:
 - correlation integral (= pair-count plot)
 - rank/frequency plot (Zipf's law)
 - CCDF (Korcak's law)

same
info

15-826 Copyright: C. Faloutsos (2009) 73

CMU SCS

Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- ➔ • More tools and **examples**
 - Discussion - putting fractals to work!
 - Conclusions – practitioner's guide
 - Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2009) 74

CMU SCS

Other applications: Internet

- How does the internet look like?


CMU

15-826 Copyright: C. Faloutsos (2009) 75

CMU SCS

Other applications: Internet

- How does the internet look like?
- Internet routers: how many neighbors within h hops?



15-826 Copyright: C. Faloutsos (2009) 76

CMU SCS

(reminder: our tool-box:)

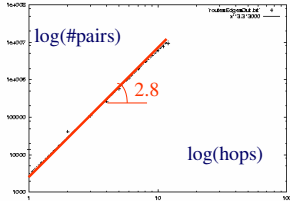
- concepts:
 - fractals, multifractals and fat fractals
- tools:
 - correlation integral (= pair-count plot)
 - rank/frequency plot (Zipf's law)
 - CCDF (Korcak's law)

15-826 Copyright: C. Faloutsos (2009) 77

CMU SCS

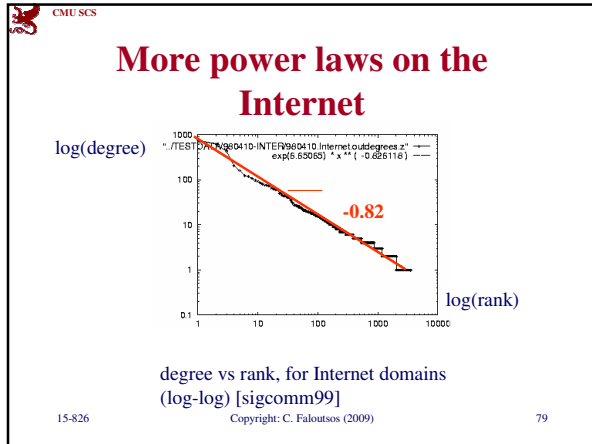
Internet topology

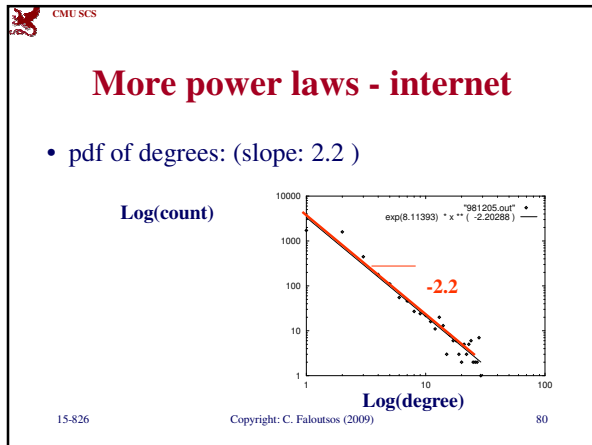
- Internet routers: how many neighbors within h hops?

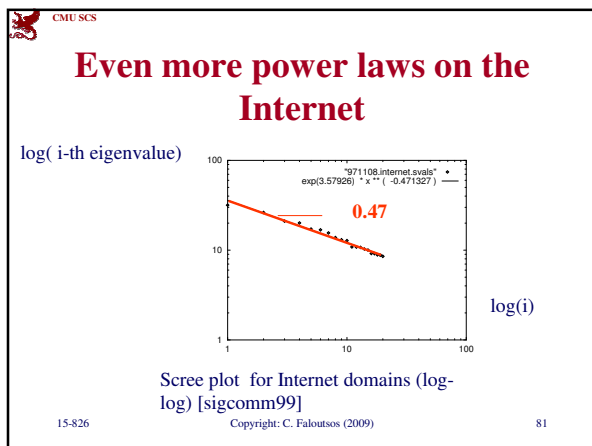


Reachability function: number of neighbors within r hops, vs r (log-log).
Mbone routers, 1995

15-826 Copyright: C. Faloutsos (2009) 78







CMU SCS

Fractals & power laws:

appear in numerous settings:

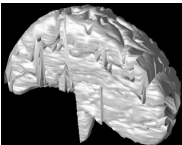
- **medical**
- geographical / geological
- social
- computer-system related

15-826 Copyright: C. Faloutsos (2009) 82

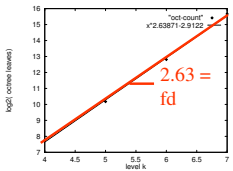
CMU SCS

More apps: Brain scans

- Oct-trees; brain-scans



Log(#octants)



octree levels

15-826 Copyright: C. Faloutsos (2009) 83

CMU SCS

More apps: Medical images


[Burdett et al, SPIE '93]:

- benign tumors: $fd \sim 2.37$
- malignant: $fd \sim 2.56$

15-826 Copyright: C. Faloutsos (2009) 84

CMU SCS

More fractals:

- cardiovascular system: 3 (!) 
- lungs: 2.9

15-826 Copyright: C. Faloutsos (2009) 85

CMU SCS

Fractals & power laws:

appear in numerous settings:

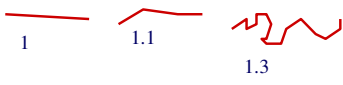
- medical
- **geographical / geological**
- social
- computer-system related

15-826 Copyright: C. Faloutsos (2009) 86

CMU SCS

More fractals:

- Coastlines: 1.2-1.58



1 1.1 1.3

15-826 Copyright: C. Faloutsos (2009) 87



More fractals:

- the fractal dimension for the Amazon river is 1.85 (Nile: 1.4)

[ems.gphys.unc.edu/nonlinear/fractals/examples.html]

A satellite image showing a dense network of green rivers and streams flowing towards a blue body of water. The image is titled '15-826' and 'Copyright: C. Faloutsos (2009)'.

More fractals:

- the fractal dimension for the Amazon river is 1.85 (Nile: 1.4)

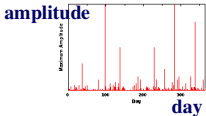
[ems.gphys.unc.edu/nonlinear/fractals/examples.html]

A black and white fractal image showing a complex, branching network of lines, representing a river system. The image is titled '15-826' and 'Copyright: C. Faloutsos (2009)'.

CMU SCS

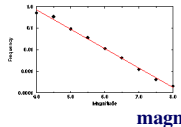
More power laws

- Energy of earthquakes (Gutenberg-Richter law) [simscience.org]



amplitude

day



log(freq)

magnitude

15-826 Copyright: C. Faloutsos (2009) 91

CMU SCS

Fractals & power laws:

appear in numerous settings:


- medical
- geographical / geological
- **social**
- computer-system related

15-826 Copyright: C. Faloutsos (2009) 92


CMU SCS

More fractals:

stock prices (LYCOS) - random walks: 1.5



1 year



2 years

15-826 Copyright: C. Faloutsos (2009) 93

CMU SCS

Even more power laws:

- Income distribution (Pareto's law)
- size of firms
- publication counts (Lotka's law)

15-826 Copyright: C. Faloutsos (2009) 94

CMU SCS

Even more power laws:

library science (Lotka's law of publication count); and citation counts:
(*citeseer.nj.nec.com* 6/2001)

log(count)

Ullman

log(#citations)

15-826 Copyright: C. Faloutsos (2009) 95

CMU SCS

Even more power laws:

- web hit counts [w/ A. Montgomery]

Web Site Traffic

log(count)

Zipf

“yahoo.com”

log(freq)

15-826 Copyright: C. Faloutsos (2009) 96

CMU SCS

Fractals & power laws:

appear in numerous settings:

- medical
- geographical / geological
- social
- **computer-system related**

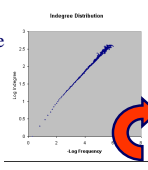
15-826 Copyright: C. Faloutsos (2009) 97

CMU SCS

Power laws, cont'd

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]

log indegree



from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]

- log(freq)

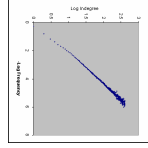
15-826 Copyright: C. Faloutsos (2009) 98

CMU SCS

Power laws, cont'd

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]

log(freq)



from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]

log indegree

15-826 Copyright: C. Faloutsos (2009) 99

CMU SCS

“Foiled by power law”

- [Broder+, WWW’00]

(log) count

in-degree (total, remote-only) distr.

“The anomalous bump at 120 on the x-axis is due a large clique formed by a single spammer”

15-826 Copyright: C. Faloutsos (2009) 100

CMU SCS

Power laws, cont’d

- In- and out-degree distribution of web sites [Barabasi], [IBM-CLEVER]
- length of file transfers [Crovella+Bestavros ‘96]
- duration of UNIX jobs [Harchol-Balter]

15-826 Copyright: C. Faloutsos (2009) 101

CMU SCS

Even more power laws:

- Distribution of UNIX file sizes
- web hit counts [Huberman]

15-826 Copyright: C. Faloutsos (2009) 102

CMU SCS

Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- ➔ Discussion - putting fractals to work!
- Conclusions – practitioner’s guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2009) 103

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]
 - ...

15-826 Copyright: C. Faloutsos (2009) 104

CMU SCS

Settings for fractals:

Points; areas (-> fat fractals), eg:

15-826 Copyright: C. Faloutsos (2009) 105

CMU SCS

Settings for fractals:

Points; areas, eg:

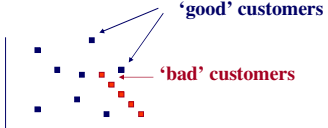
- cities/stores/hospitals, over earth's surface
- time-stamps of events (customer arrivals, packet losses, criminal actions) over time
- regions (sales areas, islands, patches of habitats) over space

15-826 Copyright: C. Faloutsos (2009) 106

CMU SCS

Settings for fractals:

- customer feature vectors (age, income, frequency of visits, amount of sales per visit)



15-826 Copyright: C. Faloutsos (2009) 107

CMU SCS

Some uses of fractals:

- Detect non-existence of rules (if points are uniform)
- Detect non-homogeneous regions (eg., legal login time-stamps may have different fd than intruders')
- Estimate number of neighbors / customers / competitors within a radius

15-826 Copyright: C. Faloutsos (2009) 108

CMU SCS

Multi-Fractals

Setting: points or objects, w/ some value, eg:

- cities w/ populations
- positions on earth and amount of gold/water/oil underneath
- product ids and sales per product
- people and their salaries
- months and count of accidents

15-826 Copyright: C. Faloutsos (2009) 109

CMU SCS

Use of multifractals:

- Estimate tape/disk accesses
 - *how many of the 100 tapes contain my 50 phonecall records?*
 - *how many days without an accident?*

The diagram shows a horizontal line representing time, labeled 'time' at the bottom. Above the line, 'Tape#1' is marked with a double-headed arrow on the left, and 'Tape# N' is marked with a double-headed arrow on the right. Three red stars are placed along the timeline: two are clustered together near the start, and one is further to the right.

15-826 Copyright: C. Faloutsos (2009) 110

CMU SCS

Use of multifractals

- how often do we exceed the threshold?

#bytes

The graph shows a vertical axis labeled '#bytes' and a horizontal axis labeled 'time'. A horizontal red line represents a threshold. A jagged line representing data fluctuates around a lower red curve labeled 'Poisson'. The data line crosses the threshold line several times.

15-826 Copyright: C. Faloutsos (2009) 111

CMU SCS

Use of multifractals cont'd

- Extrapolations for/from samples

#bytes

time

15-826 Copyright: C. Faloutsos (2009) 112

CMU SCS

Use of multifractals cont'd

- *How many distinct products account for 90% of the sales?*

20% ↗ 80%

0 0.25 0.5 0.75 1

15-826 Copyright: C. Faloutsos (2009) 113

CMU SCS

Road map

- Motivation – 3 problems / case studies
- Definition of fractals and power laws
- Solutions to posed problems
- More examples and tools
- Discussion - putting fractals to work!
- ➔ • Conclusions – practitioner's guide
- Appendix: gory details - boxcounting plots

15-826 Copyright: C. Faloutsos (2009) 114

CMU SCS

Conclusions

- Real data often **disobey** textbook assumptions (Gaussian, Poisson, uniformity, independence)

15-826 Copyright: C. Faloutsos (2009) 115

CMU SCS

Conclusions - cont'd

Self-similarity & power laws: appear in **many** cases

<p>Bad news: lead to skewed distributions (no Gaussian, Poisson, uniformity, independence, mean, variance)</p>	<p>Good news: </p> <ul style="list-style-type: none"> 'correlation integral' for separability rank/frequency plots 80-20 (multifractals) (Hurst exponent, strange attractors, renormalization theory, 116) ++)
--	---

15-826 Copyright: C. Faloutsos (2009) 116

CMU SCS

Conclusions

- tool#1: (for points) 'correlation integral':** (#pairs within $\leq r$) vs (distance r)
- tool#2: (for categorical values) rank-frequency plot** (a'la Zipf)
- tool#3: (for numerical values) CCDF:** Complementary cumulative distr. function (#of elements with value $\geq a$)

15-826 Copyright: C. Faloutsos (2009) 117

Practitioner's guide:

- tool#1:** #pairs vs distance, for a **set of objects**, with a distance function (slope = intrinsic dimensionality)

log(#pairs) internet

log(hops)

2.8

log(#pairs(within <= r))

SLOPE = 1.51847

MGcounty

1.51

log(r)

15-826 Copyright: C. Faloutsos (2009) 118

Practitioner's guide:

- tool#2:** rank-frequency plot (for **categorical attributes**)

internet domains

log(degree)

log(rank)

-0.82

Bible

log(freq)

log(rank)

15-826 Copyright: C. Faloutsos (2009) 119

Practitioner's guide:

- tool#3:** CCDF, for (skewed) **numerical attributes**, eg. areas of islands/lakes, UNIX jobs...)

log(count(>= area))

scandinavian lakes

log(area)

15-826 Copyright: C. Faloutsos (2009) 120

CMU SCS

Resources:

- Software for fractal dimension
 - <http://www.cs.cmu.edu/~christos>
 - christos@cs.cmu.edu

15-826 Copyright: C. Faloutsos (2009) 121

CMU SCS

Books

- Strongly recommended intro book:
 - Manfred Schroeder *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* W.H. Freeman and Company, 1991
- Classic book on fractals:
 - B. Mandelbrot *Fractal Geometry of Nature*, W.H. Freeman, 1977

15-826 Copyright: C. Faloutsos (2009) 122

CMU SCS

References

- [vldb95] Alberto Belussi and Christos Faloutsos, *Estimating the Selectivity of Spatial Queries Using the 'Correlation' Fractal Dimension* Proc. of VLDB, p. 299-310, 1995
- [Broder+'00] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener, *Graph structure in the web*, WWW'00
- M. Crovella and A. Bestavros, *Self similarity in World wide web traffic: Evidence and possible causes*, SIGMETRICS '96.

15-826 Copyright: C. Faloutsos (2009) 123

CMU SCS

References

- [ieeeTN94] W. E. Leland, M.S. Taqqu, W. Willinger, D.V. Wilson, *On the Self-Similar Nature of Ethernet Traffic*, IEEE Transactions on Networking, 2, 1, pp 1-15, Feb. 1994.
- [pods94] Christos Faloutsos and Ibrahim Kamel, *Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension*, PODS, Minneapolis, MN, May 24-26, 1994, pp. 4-13

15-826 Copyright: C. Faloutsos (2009) 124

CMU SCS

References

- [vlb96] Christos Faloutsos, Yossi Matias and Avi Silberschatz, *Modeling Skewed Distributions Using Multifractals and the '80-20 Law'* Conf. on Very Large Data Bases (VLDB), Bombay, India, Sept. 1996.

15-826 Copyright: C. Faloutsos (2009) 125

CMU SCS

References

- [vlb96] Christos Faloutsos and Volker Gaede *Analysis of the Z-Ordering Method Using the Hausdorff Fractal Dimension* VLD, Bombay, India, Sept. 1996
- [sigcomm99] Michalis Faloutsos, Petros Faloutsos and Christos Faloutsos, *What does the Internet look like? Empirical Laws of the Internet Topology*, SIGCOMM 1999

15-826 Copyright: C. Faloutsos (2009) 126

CMU SCS

References

- [icde99] Guido Proietti and Christos Faloutsos, *I/O complexity for range queries on region data stored using an R-tree* International Conference on Data Engineering (ICDE), Sydney, Australia, March 23-26, 1999
- [sigmod2000] Christos Faloutsos, Bernhard Seeger, Agma J. M. Traina and Caetano Traina Jr., *Spatial Join Selectivity Using Power Laws*, SIGMOD 2000

15-826 Copyright: C. Faloutsos (2009) 127

CMU SCS

References

- [Wang+02] Mengzhi Wang, Anastasia Ailamaki and Christos Faloutsos, [Capturing the spatio-temporal behavior of real traffic data](#) Performance 2002 (IFIP Int. Symp. on Computer Performance Modeling, Measurement and Evaluation), Rome, Italy, Sept. 2002

15-826 Copyright: C. Faloutsos (2009) 128

CMU SCS

Appendix - Gory details

- Bad news: There are more than one fractal dimensions
 - Minkowski fd; Hausdorff fd; Correlation fd; Information fd
- Great news:
 - they can all be computed fast!
 - they usually have nearby values

15-826 Copyright: C. Faloutsos (2009) 129

CMU SCS

Fast estimation of $fd(s)$:

- How, for the (correlation) fractal dimension?
- A: Box-counting plot:

15-826 Copyright: C. Faloutsos (2009) 130

CMU SCS

Definitions

- pi : the percentage (or count) of points in the i -th cell
- r : the side of the grid

15-826 Copyright: C. Faloutsos (2009) 131

CMU SCS

Fast estimation of $fd(s)$:

- compute $\sum(pi^2)$ for another grid side, r'

15-826 Copyright: C. Faloutsos (2009) 132

CMU SCS

Fast estimation of fd(s):

- etc; if the resulting plot has a linear part, its slope is the correlation fractal dimension D_2

log(sum(pi ^2))

log(r)

15-826 Copyright: C. Faloutsos (2009) 133

CMU SCS

Definitions (cont'd)

- Many more fractal dimensions D_q (related to Renyi entropies):

$$D_q = \frac{1}{q-1} \frac{\partial \log(\sum p_i^q)}{\partial \log(r)} \quad q \neq 1$$

$$D_1 = \frac{\partial \sum p_i \log(p_i)}{\partial \log(r)}$$

15-826 Copyright: C. Faloutsos (2009) 134

CMU SCS

Hausdorff or box-counting fd:

- Box counting plot: $\text{Log}(N (r))$ vs $\text{Log} (r)$
- r : grid side
- $N (r)$: count of non-empty cells
- (Hausdorff) fractal dimension D_0 :

$$D_0 = - \frac{\partial \log(N(r))}{\partial \log(r)}$$

15-826 Copyright: C. Faloutsos (2009) 135

CMU SCS

Definitions (cont'd)

- Hausdorff fd:

$r \quad \log(\#\text{non-empty cells})$

SLOPE = -1.5743

D_0

15-826 Copyright: C. Faloutsos (2009) 136

CMU SCS

Observations

- $q=0$: Hausdorff fractal dimension
- $q=2$: Correlation fractal dimension (**identical** to the exponent of the number of neighbors vs radius)
- $q=1$: Information fractal dimension

15-826 Copyright: C. Faloutsos (2009) 137

CMU SCS

Observations, cont'd

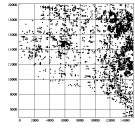
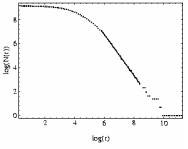
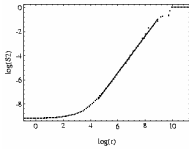
- in general, the D_q 's take similar, but not identical, values.
- except for perfectly self-similar point-sets, where $D_q=D_{q'}$ for any q, q'

15-826 Copyright: C. Faloutsos (2009) 138

CMU SCS

Examples:MG county

- Montgomery County of MD (road end-points)

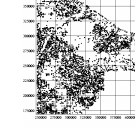
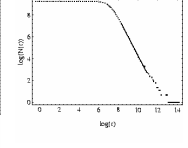
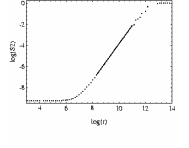




15-826 Copyright: C. Faloutsos (2009) 139

CMU SCS

Examples:LB county

- Long Beach county of CA (road end-points)

15-826 Copyright: C. Faloutsos (2009) 140

CMU SCS

Conclusions

- many fractal dimensions, with nearby values
- can be computed quickly ($O(N)$ or $O(N \log(N))$)
- (code: on the web)

15-826 Copyright: C. Faloutsos (2009) 141
